


Deep learning based on hematoxylin–eosin staining outperforms immunohistochemistry in predicting molecular subtypes of gastric adenocarcinoma

Nadine Flinner^{1,2,3,4†}, Steffen Gretser^{1†}, Alexander Quaas⁵, Katrin Bankov¹, Alexander Stoll¹, Lara E Heckmann¹, Robin S Mayer¹, Claudia Doering¹, Melanie C Demes¹, Reinhard Buettner⁵, Josef Rueschoff⁶ and Peter J Wild^{1,2,3,4,7*} 

¹ Dr. Senckenberg Institute of Pathology, University Hospital Frankfurt, Frankfurt am Main, Germany

² Frankfurt Institute for Advanced Studies (FIAS), Frankfurt am Main, Germany

³ Frankfurt Cancer Institute (FCI), Frankfurt am Main, Germany

⁴ University Cancer Center (UCT), Frankfurt am Main, Germany

⁵ Institute of Pathology, University Hospital Cologne, Cologne, Germany

⁶ Targos Molecular Pathology GmbH, Kassel, Germany

⁷ Wildlab, University Hospital Frankfurt MVZ GmbH, Frankfurt am Main, Germany

*Correspondence to: PJ Wild, Dr. Senckenberg Institute of Pathology/University Hospital Frankfurt, Theodor-Stern-Kai 7, 60596 Frankfurt am Main, Germany. E-mail: peter.wild@kgu.de

†These authors contributed equally to this study.

Abstract

In gastric cancer (GC), there are four molecular subclasses that indicate whether patients respond to chemotherapy or immunotherapy, according to the TCGA. In clinical practice, however, not every patient undergoes molecular testing. Many laboratories have used well-implemented *in situ* techniques (IHC and EBER-ISH) to determine the subclasses in their cohorts. Although multiple stains are used, we show that a staining approach is unable to correctly discriminate all subclasses. As an alternative, we trained an ensemble convolutional neuronal network using bagging that can predict the molecular subclass directly from hematoxylin–eosin histology. We also identified patients with predicted intra-tumoral heterogeneity or with features from multiple subclasses, which challenges the postulated TCGA-based decision tree for GC subtyping. In the future, deep learning may enable targeted testing for molecular subtypes and targeted therapy for a broader group of GC patients.

© 2022 The Authors. *The Journal of Pathology* published by John Wiley & Sons Ltd on behalf of The Pathological Society of Great Britain and Ireland.

Keywords: stomach neoplasms; deep learning; molecular typing; molecular diagnostic techniques; histology; computational pathology; ensemble CNN

Received 28 September 2021; Revised 4 January 2022; Accepted 31 January 2022

Conflict of interest statement: PJW has received consulting fees and honoraria (direct/institutional) for lectures from Bayer, Janssen-Cilag, Novartis, Roche, MSD, Astellas Pharma, Bristol-Myers Squibb, Hedera Dx, Thermo Fisher Scientific, Molecular Health, Sophia Genetics, Qiagen, Eli Lilly, Myriad, and Astra Zeneca. PJW has received research funding by Astra Zeneca and Thermo Fisher. JR is the co-founder and CMO of Targos Molecular Pathology GmbH (Kassel/Germany). No other conflicts of interest were declared.

Introduction

The correct subclassification of malignant tumors plays a key role in histopathological evaluation of tumor specimens. In addition to conventional histomorphology, molecular subclassification of tumor entities is becoming increasingly important for patient care. Gastric cancer (GC), one of the most lethal tumor types, can be divided into diffuse, intestinal, and mixed subclasses based on its morphologic features according to the classification of Lauren [1]. In 2014, the TCGA developed a robust molecular classification system that distinguishes four molecular subtypes of GC, which were incorporated

into the WHO classification in 2019 [2]: EBV (Epstein–Barr-virus positive, 9%), MSI (microsatellite unstable, 22%), GS (genomically stable, 19%), and CIN (chromosomally unstable, 50%) [3] with some clinical implications. Here, differences in response to chemotherapy were observed, with CIN-positive GC patients benefiting most from chemotherapy [4]. Even more interestingly, EBV and MSI subtypes appeared to respond much better to immunotherapy regimens (e.g. pembrolizumab), with overall response rates ranging from 85% to 100% [5]. Several attempts have been made to classify GC into molecular subgroups based on tissue staining techniques, without using molecular

sequencing methods: EBV cases can be identified using EBV-encoded RNA (EBER) *in situ* hybridization, and MSI can be detected using immunohistochemistry (IHC; with up to four antibodies: anti-MLH1, anti-PMS2, anti-MSH2, anti-MSH6) or fragment length analysis [5–12]. However, the distinction between GS and CIN is less clear. For these categories, decreased expression of E-cadherin [8], Lauren morphology [7], or mutational status of *TP53* [11] are used for differentiation. However, comparison of these methods with genomic data is lacking, so their trustworthiness is unknown.

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 [13]. Pseudonymized archival tissue samples were retrieved from the tissue bank of the Institute of Pathology, University Hospital Cologne (UHC) after approval by the institutional ethics board as described in the internal documents 13–091 and 10–242.

Antibodies

The following antibodies were used: anti-HER2 (clone 4B5; catalogue number 709-4493; ready-to-use dilution; Roche/Ventana, Mannheim, Germany), anti-TP53 (clone DO-7; catalogue number M7001; 1:800 dilution; DAKO, Glostrup, Denmark), anti-E-cadherin (clone NCH-38; catalogue number M3612; 1:50 dilution; DAKO), anti-MLH1 (clone M1; catalogue number 760-5091; ready-to-use dilution; Roche/Ventana), and anti-MSH2 (clone G219-1129; catalogue number 760-5093; ready-to-use dilution; Roche/Ventana).

EBV-encoded RNA (EBER) *in situ* hybridization

A fluorescein-conjugated oligonucleotide probe was used in conjunction with a monoclonal anti-fluorescein antibody (EBER-ISH probe for BOND Ready-to-Use with Ready-to-Use Anti-Fluorescein Antibody, catalogue numbers PB0589 and AR0222; Leica, Newcastle Upon Tyne, UK) and DAB as chromogen (Leica Biosystems, Wetzlar, Germany) according to the manufacturer's instructions. EBV positivity was defined as the presence of staining in more than 80% of tumor cell nuclei.

Immunohistochemistry

GC tissue sections cut at 3 μm from formalin-fixed, paraffin-embedded (FFPE) blocks were subjected to antigen retrieval (microwave oven for 10 min at 250 W; citrate buffer for TP53, EDTA for all other antibodies) and immunohistochemistry was carried out in a Benchmark immunostainer (Ventana, Tucson, AZ,

USA). Two surgical pathologists (RB and AQ) performed a blinded evaluation of the immunostained slides without knowledge of clinical data. For the analysis of p53 staining, tumors were scored in five groups based on the extent of p53 positivity by determining the frequency of tumor cells displaying strong nuclear staining (0%, 1–10%, 11–50%, 51–90%, and 91–100%). Focal weak nuclear p53 staining was regarded as background and scored as 0%. HER2 expression was scored according to the DAKO HercepTest criteria [14].

Copy number variations (CNVs)

DNA extraction and Affymetrix (Agilent) OncoScan copy number variation (CNV) analysis were performed as described previously [15].

CIN ratios

To obtain CIN ratios for the TCGA data, we downloaded the `broad_values_by_arm.txt` file from firebrowse.org for the TCGA gastric cancer project (STAD, stomach adenocarcinoma). Arms with a value above 0.1 and below -0.1 were counted as altered. The CIN ratio was calculated as modified arms divided by total arms according to Kohlruss *et al* [16]. To make the TCGA data comparable with OncoScan results, only arms covered by OncoScan CNV arrays were included.

Tumor annotation and patient cohorts

For training, validation, and internal testing 133 whole-slide images (WSIs) (27 EBV, 38 MSI, 32 GS, 36 CIN) from the TCGA-STAD project were used. The 65 UKC images were used exclusively for external testing. In all histological images of GC, the cancerous regions were outlined by an expert pathologist (SG), excluding regions of low image quality (e.g. pen markings). The slides were visualized using QuPath [17], and the annotations were made using the tools available there (polygon, wand, brush). For each WSI, up to 100 image patches of 600×600 pixels representing an area of $300 \times 300 \mu\text{m}$ were extracted from these cancerous regions using the open slide library (v1.1.1) [18] for WSI handling and skImage (v0.16.2) [19] for scaling. For convolutional neuronal network (cNN) training and evaluation, we used the molecular subclass as label. For the TCGA cohort, we took that label directly from the supplementary material, Table S11.1a of the original publication [3]. For the UKC cohort, our ground truth label used during testing was determined using the staining approach described herein.

Neural network models and class detection

The individual cNNs classifying the molecular subtypes were trained using the pyTorch [20] (torch: v1.0.1; torchvision: 0.2.2) and the fastai [21] (v1.0.52) library. A (on ImageNet pretrained) DenseNet161 architecture [22] was used, and the weights of the convolutional layers were kept fixed, the classification layer consisted

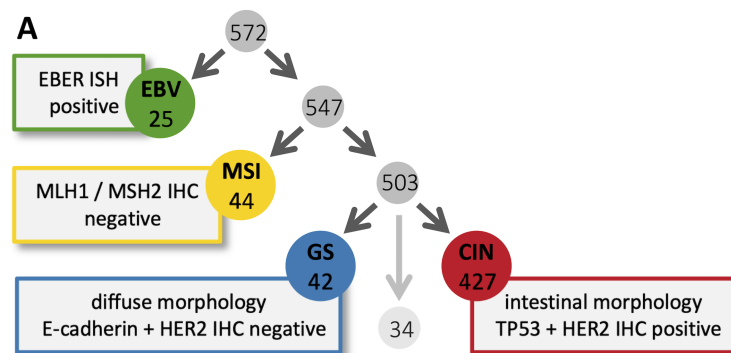
of three fully connected layers, and a learning rate of $1e-2$ was used. Images were augmented slightly in every epoch (flipping and rotating by 90°). Networks were trained for seven epochs using 244×244 pixels from each patch with a weight decay of 0.1, a dropout of 0.4, and a batch size of 40. The network with the lowest error rate at the end of the epochs was saved for further use.

Because each individual cNN was trained in a balanced manner, and due to the low number of EBV-positive cases in the TCGA cohort in every iteration, 21 WSIs from each subclass (in total 84) were used for training, four for validation (in total 16) and the three or four for the internal/hold-out test (three for EBV and four for CIN, GS and MSI; in total 15). From each WSI 30 patches were used, in case less were available (e.g. because of a too-small tumor area), the other WSIs were oversampled to balance the dataset. Oversampling the less frequent classes does not strongly decrease error rates, while oversampling a limited number of EBV cases (e.g. use only ten EBV cases for training but 21 for MSI, GS, and CIN) increases error rates.

For the ensemble, the individual cNNs were trained using a bagging-like approach: The patients were randomly assigned to the datasets and for each patient, 30 image patches from the WSIs were randomly used so that each cNN was trained with slightly different patients and image patches in the training dataset. The final ensemble prediction for one image patch is a simple consensus/majority vote over all individual cNNs. In addition, a patient label was determined by simple consensus/majority vote over image patches.

Statistics

Statistical tests were performed with the `scipy.stats` package [23] in Python as implemented in the '`f_one-way`', '`ks_2samp`', '`ttest_ind`' method. For multiple testing, the '`multicomp.allpairtest`' method was used with method '`bonf`' for *P* value adjustment.



Results and discussion

Design and validation of a staining approach

We decided to classify patients from our own GC cohort collected at UKC into the different molecular subclasses using an approach integrating the previously used criteria (Figure 1A). In total, 538 of 572 cases were analyzable (94%). EBV patients were detected using EBER *in situ* hybridization (25/538, 5%). Patients with negative immunoreactivity for MLH1 or MSH2 were classified as MSI-positive (44/538, 8%). Compared with the TCGA cohort, these two groups were underrepresented in our German cohort. The remaining patients were separated into GS (42/538, 8%) and CIN (427/538, 79%) based on Lauren morphology, loss of E-cadherin expression, and TP53 overexpression.

In the original TCGA publication, the latter distinction was based on clustering of an SNP array-based copy number analysis [3]. However, a measure with a clear threshold value such as the CIN ratio [16] would be more suitable for clinical routine. To now verify whether our staining approach can reproduce the separation between GS and CIN, we additionally performed molecular analyses of the copy number changes using OncoScan CNV arrays and computed CIN ratios, covering all subclasses of the staining approach (Figure 1B). Supplementary material, Figure S1 provides gains and losses per chromosome for all patients from the Cologne cohort for whom CNV data could be generated. For example, cases #2 and #8 from the Cologne cohort classified as CIN by staining methods nicely show the discrepancy with the OncoScan CNV data; both cases show hardly any aberrations (supplementary material, Figure S1). Conversely, numerous GC samples labeled as GS by staining techniques show numerous aberrations (e.g. #37-41, #43, and #46). The calculation of the CIN ratio [16] thus represents a simple, quantitative, and easily interpretable measure for the estimation of chromosomal instability (CIN), which enables the differentiation of GS and CIN.

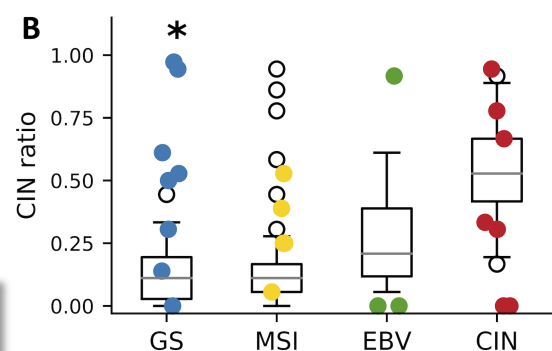


Figure 1. Gastric cancer cohort and CIN ratios determined by *in situ* hybridization. (A) Schematic representation of the staining approach to determine the molecular subtypes in the UKC cohort. The numbers in the circles indicate the number of patients in the UKC cohort assigned to each class. Thirty-four patients could not be classified using this approach. (B) Boxplot of CIN ratios separated by molecular subclasses for the TCGA dataset. CIN ratios for selected cases from the UKC dataset are shown in color for each subclass as determined by the staining approach. There were 6/8 GS cases that had abnormally high CIN ratios compared with the TCGA data. An asterisk marks classes where TCGA and UKC data are from different distributions (*ks* test with $p < 0.05$).

Although there was no defined threshold to distinguish CIN from GS cases in the TCGA cohort, as expected, CIN cases had a higher CIN ratio compared with the other classes in the TCGA cohort [Figure 1B, mean CIN ratio: CIN 0.53, EBV 0.26, MSI 0.15, GS 0.12; ANOVA P value = $3.9\text{e-}52$ and Kolmogorov–Smirnov (k_s) test $p < 0.05$ (with Bonferroni correction) for CIN versus all other and EBV versus GS]. CIN ratios of GS cases based on *in situ*-based subtyping were significantly different from GS cases within TCGA (k_s test $p < 0.05$). Thus, GS labeling based on the two key features defined in TCGA (diffuse histology and altered cell adhesion) [3] is insufficient for correct identification of GS cases and cannot replace the explicit determination of copy number alterations. Moreover, this observation is consistent with the fact that expression of cell adhesion proteins (e.g. B1/B3 integrins) was increased [3] in GS, and detection of E-cadherin loss by IHC [24] is not an appropriate criterion to identify GS cases. For the classes GS and CIN, no gene with sufficient differential expression was found in the TCGA data, so improved staining-based classification appeared difficult (Supplementary materials and methods and Figure S2).

Deep learning as an alternative approach

Overall, separation of GC patients was not possible using simple staining methods. A possible alternative to determine the correct molecular subtype without comprehensive molecular profiling could be deep learning, which is gaining attention in some medical data analysis tasks [25]. For example, using deep learning, cancer patient survival and mutation status in tumors could be predicted from H&E images of tumors of the lung [26], prostate [27,28], brain [29], and other locations. Recently, molecular subclasses based on tissue morphology were also shown to be successfully predicted using convolutional neuronal networks (cNNs) in bladder [30] and colorectal [31] cancer. In addition, Kather *et al* showed that MSI can be correctly predicted in gastric and colorectal cancer using artificial intelligence [32,33].

Therefore, we decided to train a classifier (Figure 2A) to predict the molecular subtype of GC based solely on H&E histomorphology using expert annotated WSIs from TCGA as a study cohort. Data from TCGA were used for training and model selection (here called validation), as well as for internal testing. The UKC cohort was used exclusively for external approval (here called testing) and not for training. The training, validation, and testing datasets were balanced for the image patches. Each patient belonged exclusively to only one dataset. We found that training was limited by the subclass with the lowest case number (EBV) and that oversampling this class to include more CIN, MSI, and GS patients did not increase performance (Supplementary materials and methods, Figure S3, and Table S1). Ideal parameters were achieved via a grid search (Supplementary materials and methods and Figure S4). An

individual cNN achieved a mean validation error rate of $43 \pm 7\%$ and a slightly higher mean test error of $49 \pm 8\%$ for TCGA data, while there were no significant differences between the two error rates (supplementary material, Figure S5). Next, we wanted to test whether the trained cNN also worked on the independent UKC test dataset. As expected, the test error rate was significantly increased to $62 \pm 6\%$ (Figure 2B and supplementary material, Figure S5 and Table S2).

To test whether the GS cases, likely to be misclassified by our staining approach, were responsible for this observation, we removed them from the test datasets. And in these modified test datasets, the resulting error rates were no longer significantly different (Figure 2B). In line with this, the number of potentially detected GS cases in the full UKC test dataset was $11 \pm 7\%$, so it was below random guessing (25% for four classes) and differed from the other classes, where the number of true positives was between 40% (CIN) and 58% (EBV) (supplementary material, Table S2). In contrast, the true positive rate in the hold-out TCGA test dataset was in the same range for all four classes (MSI, 43%; GS, 55%; supplementary material, Table S2). To further support the hypothesis that the GS cases caused the problem, we additionally trained cNNs with three classes. In each experiment, we used only a subset of the molecular subtypes and removed one class from all datasets (training, validation, and hold-out test data). And, indeed, all experiments where the GS class was included in the training data showed significant differences between the error rates of the TCGA and UKC hold-out test datasets. Interestingly, no such differences between the two test error rates were observed in the experiment where the GS class was excluded (supplementary material, Figure S6). Taken together, this strengthens the hypothesis that the GS class labels determined via morphology and staining techniques contain errors and that our models trained on TCGA data are generalizable to the external UKC test data. But it is important to mention that it is still possible that the model does not generalize for the GS class, since we cannot test for this fact: a proper relabeling of the UKC cohort was still not possible, for example, because MSI detection using IHC only achieves $\sim 95\%$ accuracy [12]. In addition, a clear separation of CIN and GS cases in the TCGA dataset based on the CIN ratio was not possible due to the lack of a unique cutoff (Figure 1B). Therefore, we decided to continue our experiments with the UKC(–GS) test dataset. Remarkably, however, this potential mislabeling of diffuse cases also showed that our model, which can correctly predict GS cases in the TCGA test dataset, not only focuses on diffuse morphology but also learned true features of the GS subtype.

Kather *et al* calculated a consensus across all tiles in a GC patient to determine a patient-based classification [33]. For an independent test dataset (KCCH Gastric), they achieved an AUC (area under the curve) of 0.69 on patient level for a two-class model (MSI-high versus MSS) in GC [33]. Similarly, we achieved better performance when considering patient-based consensus

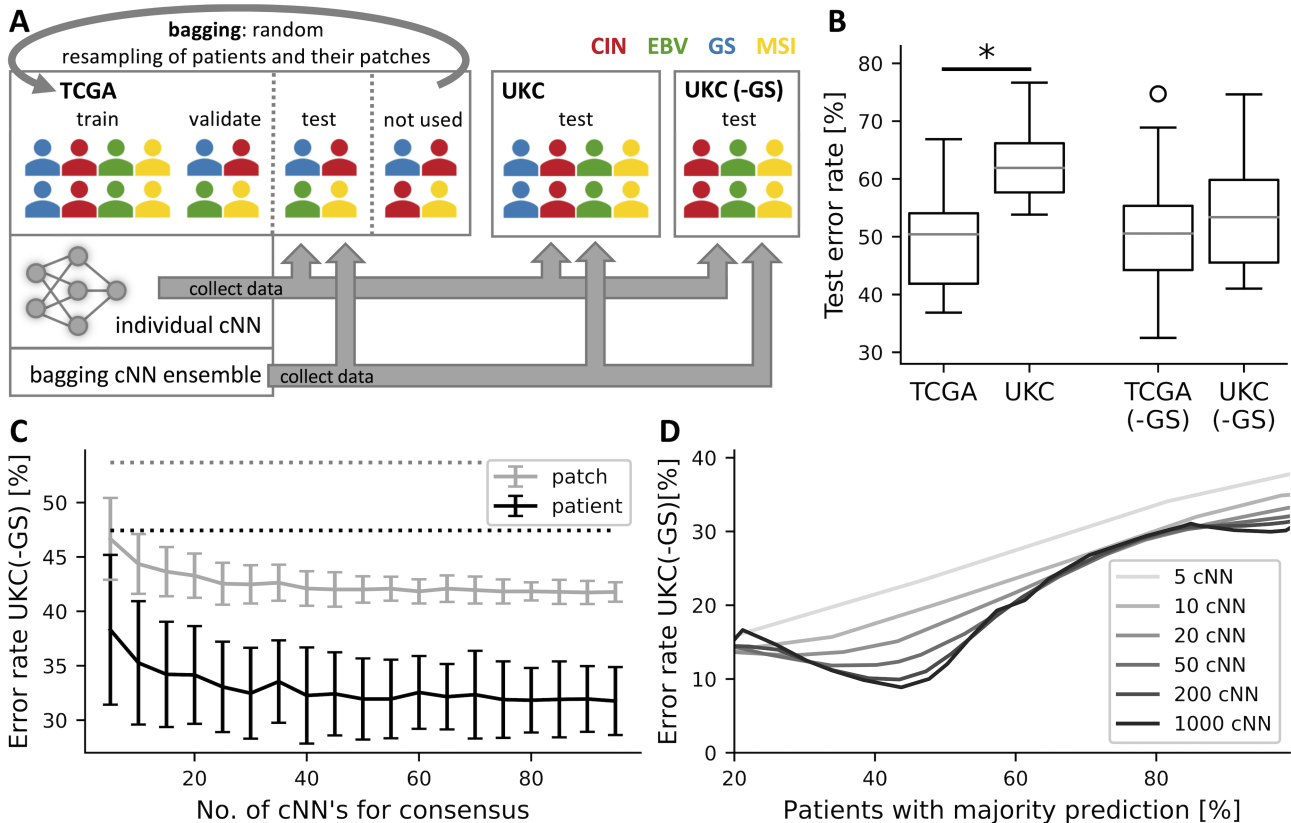


Figure 2. Study design, and single and ensemble cNN performance. (A) Study design of the training procedure. Individual DenseNet161 cNNs were trained with TCGA data to distinguish between the four molecular subclasses. Each model was trained using the bagging approach. The TCGA dataset was split into a training, a validation (for model selection), and a test dataset; the remaining patients were not used in this round. Later, multiple individual cNNs were combined into an ensemble and a consensus prediction was made for a single image patch. The individual and ensemble cNNs were then used for predictions on unseen test data. (B) Performance of single cNNs for the TCGA and UKC cohorts. The error rates for test data at the image patch level are shown graphically. No significant difference was observed for TCGA and UKC datasets without the GS class (-GS) (t -test $p < 0.05$). (C, D) Performance of ensemble cNNs for the UKC cohort. To obtain the prediction for a patient, the most frequently selected class of a cNN (four-class model) was chosen. To obtain consensus prediction for each image/patient, the most frequently selected class determined by multiple cNNs was accepted. (C) The number of cNNs used to calculate the consensus affected the error rate for the hold-out UKC test dataset (excluding GS cases). Dotted lines give the mean error rates for the prediction using a single cNN. (D) Under the condition that a certain number of cNNs have the same result to be counted, we could not make a prediction for every patient. For those patients for whom a prediction could be made under the above condition, the error rates decreased significantly, while the remaining patients were excluded. For example, if we accept a consensus prediction for only 40% of all patients and use an ensemble with 200 individual cNNs, we observe an error rate of only ~10%. For more details on this plot, please refer to supplementary material, Figure S7.

(patch-based: 53.5% versus patient-based: 47.4%) and achieved an AUC of 0.76 with our four-class model on our independent UKC(-GS) test dataset on the patient level. A recent multi-center study, including many more patients in training compared with our study (90 EBV-positive; 182 MSI-positive), trained several two-class cNNs to detect either EBV or MSI in GC and reached AUCs of 0.81 and 0.76 for an external test cohort [34]. Still, our AUC values are in a similar range and in line with their findings. Of note, our four-class model also performed better in detecting EBV compared with MSI for the external UKC cohort (supplementary material, Table S2).

In addition to being able to perform majority voting for a patient's individual tiles, one can also perform majority voting on multiple cNNs, a so-called ensemble cNN, which we trained using bagging to obtain slightly

different individual cNNs (Figure 2A). For bagging, the training and validation datasets coming from the TCGA data differ for each individual cNN: we randomly assign patients to one of the datasets and then randomly draw 30 image patches from its pool of 100 pre-extracted patches. Due to the overrepresentation of MSI, GS, and CIN, there were patients who were not used during the corresponding round. But these patients can be used in one of the other rounds. A large decrease was observed for both patch-based and patient-based error rates in the external UKC(-GS) test data, even when only five cNNs were included in the ensemble cNN (patient-based five cNNs: 35.6%). Using even more cNNs (up to ~40 cNNs) in the ensemble further decreased the patient-based error rate to ~32% (Figure 2C). In addition, ensemble cNNs trained using the bagging approach always perform better compared with vanilla ensembles

(ensembles trained with the same datasets but different random seeds; supplementary material, Table S3).

Prediction using multiple cNNs, each trained with a slightly different but overlapping dataset, allowed us to make predictions only when a certain number of networks reached the same result as an additional quality criterion. To do this, we enforced that a certain percentage of the ensemble supported the majority vote and recorded the error rate for the subset of patients for whom enough individual cNNs supported the decision. Assuming that well-supported consensus prediction is only possible for half of all patients, the error rate can be reduced to 15% if 50 cNNs are used for majority voting. Using this method, it was even possible to reduce the error rate to below 10% (Figure 2D and supplementary material, Figure S7) while estimating not only MSI [33] but all four subtypes.

To better understand which morphological features are important for the prediction of ensemble cNNs, an examination of the confidently predicted image patches revealed that EBV is associated with a high density of lymphocytes, MSI with duct-like structures, GS with diffuse morphologies, and CIN with necrotic areas (supplementary material, Figure S8).

Detailed comparison of deep learning and staining-based labels

As mentioned previously, selected GC cases were analyzed using OncoScan copy number variation (CNV) arrays (Affymetrix), which also provide reliable labels for the UKC dataset. The selection of samples for this ground truth generation is detailed in supplementary material, Figure S9. In general, we noticed that samples with high homogeneity in the predicted class of all tiles were those with a valid consensus prediction across multiple cNNs (Figure 3A and supplementary material, Figure S9). Interestingly, there were also potential GC patients in the UKC dataset with high predicted homogeneity and incorrect consensus prediction according to the *in situ*-based label: for example, almost all tiles of case #2 (Figure 3A) were predicted to be positive for EBV subtype. However, the label assigned by the staining method was CIN. In general, EBER-ISH can lead to false negatives [35,36] and together with a measured CIN ratio of 0 (supplementary material, Table S4), which clearly contradicts the staining-based label, it is possible that the safe consensus prediction is not an error of the cNN but again one of the staining-based method. In addition, case #20 was predicted to be CIN by the ensemble cNN, although the EBER-ISH was positive. However, because the CIN ratio of this case was quite high (0.81 versus 0.61, the maximum observed value from TCGA), its actual classification was also controversial. For the remaining three cases with incorrect but confident prediction, we either had no tissue left for CNV analysis or the proportion of aberrant cells was too low for correct analysis using CNV arrays. Also, on a more global level (of course limited to our 22 cases with available CNV array data),

the cNN approach was better at predicting CIN-high (6/10) and CIN-low (9/12) compared with the staining approach, which correctly assigned only 3/10 and 8/12 cases, respectively (Figure 3B and supplementary material, Table S4). According to Kohlruss *et al* [16], cases are considered as CIN-high if their CIN ratio is above 0.5, while also thresholds in the range of 0.4–0.6 would come to the same conclusion. These results again highlighted the error-proneness of our staining-based method and further demonstrated that determining GC molecular subtypes using *in situ* hybridization was not feasible, whereas deep learning was a more powerful tool.

Gastric cancer (GC) intra-tumor heterogeneity

In general, it was difficult to assess whether there was true intra-tumor heterogeneity in the remaining cases with respect to molecular subclasses, while the error rate for a single tile was over 40% on the external UKC(–GS) test data. To test whether there are patients with true intra-tumor heterogeneity, we tested whether we could identify patients in which the different subtypes are localized in different spatial regions. To do this, we calculated the average pairwise distance between all tiles of a subclass and tested whether this distance was below the 95% confidence interval for a randomly drawn set of tiles of the same size. Indeed, in 27/53 patients (50.1%, Figure 3C) of the UKC(–GS) test dataset, at least one of the molecular subclasses was confined to a specific area of tissue, whereas the most common subclass in these cases usually accounted for only 40–90% of all tiles (24/27 cases) and true tumor heterogeneity could be suspected. An example where two subtypes were detected that were in different regions of the tumor is shown in Figure 3D (see supplementary material, Figure S10 for larger image patches of both regions). In this case, the consensus prediction (EBV) is false, but a large area is also classified as MSI, which is consistent with ground truth. Such tumor heterogeneities can also be observed in the TCGA dataset (supplementary material, Figure S11) for 54/134 patients (40.3%); here, we also trained ensemble cNNs but only collected data if the corresponding patient was in the hold-out test dataset in the respective run. Consistent with our data, intra-tumor heterogeneity in GC is well known: For example, the Lauren classification already accounts for the mixed type, which has different histomorphologic growth patterns within a tumor [1] but is thought to have a clonal origin. The HER2 status of a tumor also varies at different sites [37], so it is also possible that different molecular subtypes exist within a tumor. In addition to our data suggesting that heterogeneity of molecular subtypes according to TCGA definition is a common phenomenon, there are already case reports in which intra-tumoral heterogeneity has been observed for EBV [36,38] or MSI [39,40]. Although GC heterogeneity is known, its detection is still challenging in the laboratory and deep learning could help. In general,

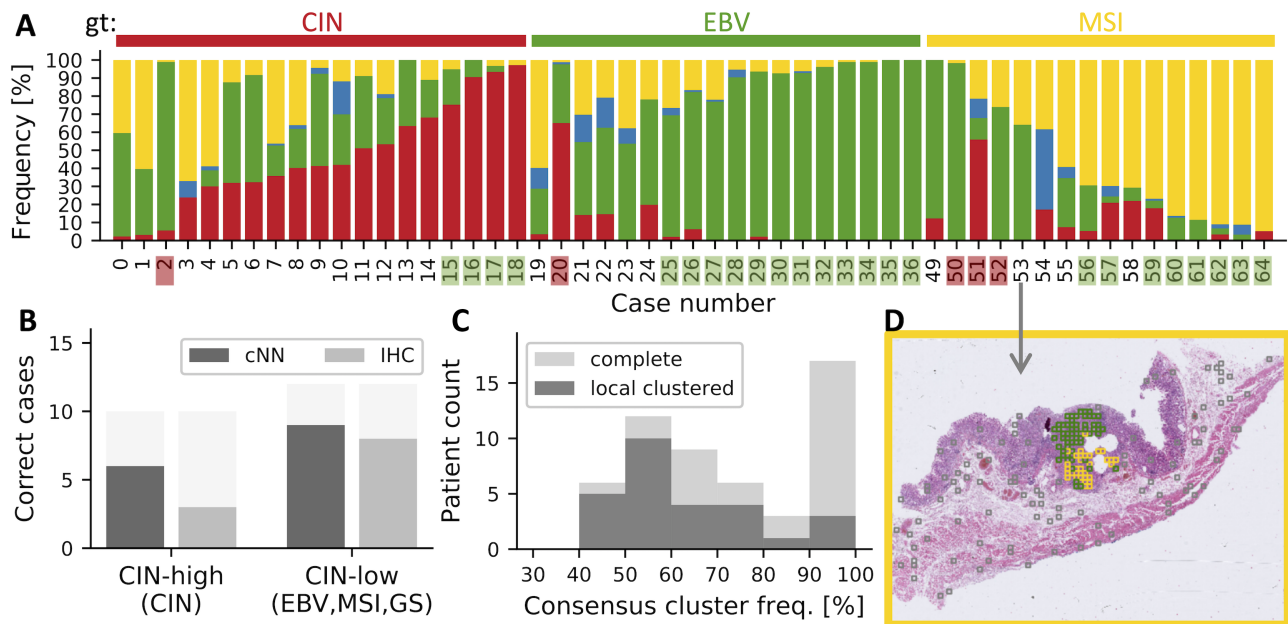


Figure 3. Gastric cancer heterogeneity and CIN prediction of the UKC cohort based on deep learning from H&E morphology. (A) Four class ensemble cNN models trained with TCGA data. The frequency of consensus prediction for all tiles of individual patients [UKC(-GS) test data] is shown. The ground truth (gt) is indicated in the top row. Green/red squares below indicate patients with secure correct/incorrect consensus prediction using ensemble cNNs. (B) Comparison of ensemble cNN-based deep learning using the staining approach in predicting CIN-high (molecular subclass designation: CIN) and CIN-low (molecular subclass designation: EBV, GS, MSI). CIN ratios above 0.5 are usually considered chromosomally unstable. (C) Histogram for the frequencies of the class with the consensus prediction (whether correct or not) for the UKC(-GS) test dataset (light grey). Dark grey indicates the number of patients in whom at least one molecular subclass was found at a given location and is counted as locally clustered. (D) Example of a whole-slide image for patient #53 where more than one subclass was found, each at a defined region.

heterogeneity could be of enormous importance as data of its influence on the therapy success of, for example, immune checkpoint inhibition is missing.

The observation of true tumor heterogeneity in many patients indicated that there are image tiles labeled with a false ground truth, since only one global label is available for each patient. For this reason, we started a second round of training and first generated a TCGA core dataset in which we included only images that were correctly classified by consensus prediction while the patient was in the hold-out test dataset (supplementary material, Figure S12). Using this core TCGA dataset for training, we obtained lower error rates for individual tiles as well as for complete patients. However, for consensus prediction with an ensemble cNN, the error rates in the external UKC(-GS) test data did not decrease as much as for the complete dataset. This indicated that removing potentially mislabeled data indeed reduces error rates; however, it seemed possible that images with, for example, atypical features for the respective subclass were removed when creating the core dataset. Thus, in the future, it would be important, to first experimentally identify patients with tumor heterogeneity, which is especially true for the test dataset, because only with this information can the correct error rates be determined, and each mislabeled image reduces the observed accuracy of the learned model. Second, more sophisticated label de-noising methods should be tested in training.

Molecular classification of gastric cancer (GC)

It needs to be validated whether the postulated decision tree is a valid method to classify all GC patients, because patients with multiple characteristics exist (e.g. patient #20 or #53 of the UKC cohort; Figure 3 and supplementary material, Table S4). Moreover, the data-rich molecular clustering in the TCGA publication and the final decision tree partially contradict each other in up to 26% of all patients (Supplementary materials and methods). Thus, comparing our error rate of the cNN ensemble (~33%) with that of the full clustering (22–26%), we can see that we are already close to the optimum. Moreover, it is also unclear which molecular classification system is most appropriate for GC as other systems exist – for example, the ACRG (Asian Cancer Research Group) system based on mRNA expression [41], systems based on methylation data [42–44], or systems that attempt to combine ACRG and TCGA [8,9]. This suggests that there is still a long way to go to perfectly classify GC into molecular subtypes [45].

Conclusions

The simplified molecular TCGA subclasses could be predicted by deep learning directly based on H&E staining, and no systematic errors were included, as it was with the easy-to-use staining approach based on WHO

rules. Those morphologic and staining methods for sub-classification of GCs were not adequate to distinguish all four molecular subtypes. Due to the inaccurate classification (especially of GS/CIN) as well as the overlap of the four GC subgroups, the artificial intelligence (AI)-based model was less trainable: with each mislabeled image, the observed accuracy of the learned model decreased. Possible causes of mislabeling in the TCGA cohort included simplified classification using a decision tree, the fact that certain tumors belong to more than one class, and intra-tumor heterogeneity. Therefore, it is promising to further refine this cNN approach in the future with additional training data (especially important for the rarer EBV class) to reduce error rates.

Acknowledgements

The results are partly based on data generated by the TCGA Research Network (<http://cancergenome.nih.gov>). NF was funded by the Mildred-Scheel-Career Center (MSNZ) Frankfurt and the Alfons und Gertrud Kassel Foundation. We thank the team of the Senckenberg Biobank [part of the interdisciplinary biobank and database (iBDF) Frankfurt] for excellent technical assistance and active support.

Author contributions statement

NF, SG and PJW conceptualized and designed the study. CD preprocessed the data. AQ, KB, AS, LEH, RSM, MCD, RB and JR contributed tumor samples and corresponding molecular and clinical data. SG, PJW, AQ, JR and RB provided histopathology expertise. NF, SG, KB, MCD and CD performed the data analysis. PJW and SG verified the underlying data. All the authors contributed to interpretation of the results. NF, SG, CD, AQ and PJW wrote the first draft of the manuscript and all the authors critically revised the manuscript. All the authors approved the final version of the manuscript. All the authors decided to submit this study and agreed to be accountable for all aspects of the work as recommended by the International Committee of Medical Journals Editors (ICMJE) authorship criteria.

Data availability statement

Generated data within this study are available upon reasonable request from the authors. Source code for training and testing of the bagging ensemble cNNs is included in supplementary material, Code S1.

References

1. Lauren P. The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma. An attempt at a histological classification. *Acta Pathol Microbiol Scand* 1965; **64**: 31–49.
2. Fukayama M, Goldblum JR, Miettinen M, *et al.* *Digestive System Tumours, WHO Classification of Tumours* (5th edn). International Agency for Research on Cancer: Lyon, 2019.
3. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 2014; **513**: 202–209.
4. Sohn BH, Hwang JE, Jang HJ, *et al.* Clinical significance of four molecular subtypes of gastric cancer identified by the Cancer Genome Atlas Project. *Clin Cancer Res* 2017; **23**: 4441–4449.
5. Kim HS, Shin SJ, Boem SH, *et al.* Comprehensive expression profiles of gastric cancer molecular subtypes by immunohistochemistry: implications for individualized therapy. *Oncotarget* 2016; **7**: 44608–44620.
6. Martinez-Ciarpaglini C, Fleitas-Kanonnikoff T, Gambardella V, *et al.* Assessing molecular subtypes of gastric cancer: microsatellite unstable and Epstein–Barr virus subtypes. Methods for detection and clinical and pathological implications. *ESMO Open* 2019; **4**: e002470.
7. Yoon J, Sy K, Brezden-Masley C, *et al.* Histo- and immunohistochemistry-based estimation of the TCGA and ACRG molecular subtypes for gastric carcinoma and their prognostic significance: a single-institution study. *PLoS One* 2019; **14**: e0224812.
8. Ahn S, Lee SJ, Kim Y, *et al.* High-throughput protein and mRNA expression-based classification of gastric cancers can identify clinically distinct subtypes, concordant with recent molecular classifications. *Am J Surg Pathol* 2017; **41**: 106–115.
9. Setia N, Agoston AT, Han HS, *et al.* A protein and mRNA expression-based classification of gastric cancer. *Mod Pathol* 2016; **29**: 772–784.
10. Koh J, Ock CY, Kim JW, *et al.* Clinicopathologic implications of immune classification by PD-L1 expression and CD8-positive tumor-infiltrating lymphocytes in stage II and III gastric cancer patients. *Oncotarget* 2017; **8**: 26356–26367.
11. Daun T, Nienhold R, Paasinen-Sohns A, *et al.* Combined simplified molecular classification of gastric adenocarcinoma, enhanced by lymph node status: an integrative approach. *Cancers (Basel)* 2021; **13**: 3722.
12. Bae YS, Kim H, Noh SH, *et al.* Usefulness of immunohistochemistry for microsatellite instability screening in gastric cancer. *Gut Liver* 2015; **9**: 629–635.
13. Council for International Organizations of Medical Sciences. International ethical guidelines for biomedical research involving human subjects. *Bull Med Ethics* 2002; **182**: 17–23.
14. Rüschoff J, Hanna W, Bilous M, *et al.* HER2 testing in gastric cancer: a practical approach. *Mod Pathol* 2012; **25**: 637–650.
15. Oehl K, Vrugt B, Wagner U, *et al.* Alterations in BAP1 are associated with cisplatin resistance through inhibition of apoptosis in malignant pleural mesothelioma. *Clin Cancer Res* 2021; **27**: 2277–2291.
16. Kohlruß M, Reiche M, Jesinghaus M, *et al.* A microsatellite based multiplex PCR method for the detection of chromosomal instability in gastric cancer. *Sci Rep* 2018; **8**: 12551.
17. Bankhead P, Loughrey MB, Fernández JA, *et al.* QuPath: open source software for digital pathology image analysis. *Sci Rep* 2017; **7**: 16878.
18. Goode A, Gilbert B, Harkes J, *et al.* OpenSlide: a vendor-neutral software foundation for digital pathology. *J Pathol Inform* 2013; **4**: 27.
19. van der Walt S, Schönberger JL, Nunez-Iglesias J, *et al.* scikit-image: image processing in Python. *PeerJ* 2014; **2**: e453.
20. Paszke A, Gross S, Massa F, *et al.* PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* (Vol. **32**), 2019. <https://doi.org/10.48550/arXiv.1912.01703>

21. Howard J, Gugger S. fastai: a layered API for deep learning. *Information* 2020; **11**: 108.
22. Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks. *Proceedings – 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, January 2017; 2261–2269.
23. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020; **17**: 261–272.
24. Ascaño JJ, Frierson H Jr, Moskaluk CA, et al. Inactivation of the E-cadherin gene in sporadic diffuse-type gastric cancer. *Mod Pathol* 2001; **14**: 942–949.
25. Norgeot B, Glicksberg BS, Butte AJ. A call for deep-learning healthcare. *Nat Med* 2019; **25**: 14–15.
26. Coudray N, Moreira AL, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *bioRxiv* 2017; 197574. [Not peer reviewed].
27. Arvaniti E, Fricker KS, Moret M, et al. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Sci Rep* 2018; **8**: 12054.
28. Bulten W, Pinckaers H, van Boven H, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol* 2020; **21**: 233–241.
29. Mobadersany P, Yousefi S, Amgad M, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc Natl Acad Sci U S A* 2018; **115**: E2970–E2979.
30. Woerl AC, Eckstein M, Geiger J, et al. Deep learning predicts molecular subtype of muscle-invasive bladder cancer from conventional histopathological slides. *Eur Urol* 2020; **78**: 256–264.
31. Sirinukunwattana K, Domingo E, Richman SD, et al. Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning. *Gut* 2021; **70**: 544–554.
32. Ehle A, Grabsch HI, Quirke P, et al. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. *Gastroenterology* 2020; **159**: 1406–1416.e11.
33. Kather J, Pearson AT, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med* 2019; **25**: 1054–1056.
34. Muti HS, Heij LR, Keller G, et al. Development and validation of deep learning classifiers to detect Epstein–Barr virus and microsatellite instability status in gastric cancer: a retrospective multicentre cohort study. *Lancet Digit Health* 2021; **3**: e654–e664.
35. Alsina M, Gullo I, Carneiro F. Intratumoral heterogeneity in gastric cancer: a new challenge to face. *Ann Oncol* 2017; **28**: 912–913.
36. Böger C, Krüger S, Behrens HM, et al. Epstein–Barr virus-associated gastric cancer reveals intratumoral heterogeneity of *PIK3CA* mutations. *Ann Oncol* 2017; **28**: 1005–1024.
37. Gullo I, Carneiro F, Oliveira C, et al. Heterogeneity in gastric cancer: from pure morphology to molecular classifications. *Pathobiology* 2018; **85**: 50–63.
38. Miyabe K, Saito M, Koyama K, et al. Collision of Epstein–Barr virus-positive and -negative gastric cancer, diagnosed by molecular analysis: a case report. *BMC Gastroenterol* 2021; **21**: 97.
39. Mathiak M, Warneke VS, Behrens HM, et al. Clinicopathologic characteristics of microsatellite instable gastric carcinomas revisited: urgent need for standardization. *Appl Immunohistochem Mol Morphol* 2017; **25**: 12–24.
40. Ottini L, Palli D, Falchetti M, et al. Microsatellite instability in gastric cancer is associated with tumor location and family history in a high-risk population from Tuscany. *Cancer Res* 1997; **57**: 4523–4529.
41. Cristescu R, Lee J, Nebozhyn M, et al. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. *Nat Med* 2015; **21**: 449–456.
42. Li T, Chen X, Gu M, et al. Identification of the subtypes of gastric cancer based on DNA methylation and the prediction of prognosis. *Clin Epigenetics* 2020; **12**: 161.
43. Lian Q, Wang B, Fan L, et al. DNA methylation data-based molecular subtype classification and prediction in patients with gastric cancer. *Cancer Cell Int* 2020; **20**: 349.
44. Yoo S, Chen Q, Wang L, et al. Molecular and cellular heterogeneity of gastric cancer explained by methylation-driven key regulators. *bioRxiv* 2020; 2020.01.27.920744. [Not peer reviewed].
45. Wang Q, Liu G, Hu C. Molecular classification of gastric adenocarcinoma. *Gastroenterology* 2019; **12**: 275–282.

SUPPLEMENTARY MATERIAL ONLINE

Supplementary materials and methods

Figure S1. Copy number variation profiles

Figure S2. mRNA expression for *ERBB2*, *CDH1*, *ITGB1*, and *TP53* in the TCGA gastric adenocarcinoma cohort

Figure S3. Training datasets should be balanced in terms of the number of patients

Figure S4. Parameter screening

Figure S5. Mean error rates on patch level

Figure S6. Test error rates for three-class models

Figure S7. Relation of error rate and the number of patients with a valid consensus prediction under different thresholds

Figure S8. Securely predicted image patches

Figure S9. Heterogeneity in the UKC dataset

Figure S10. Randomly selected tiles from Figure 3D

Figure S11. Heterogeneity in the TCGA dataset

Figure S12. Noisy labels in cNN training

Table S1. Required oversampling for different subclasses and experiments

Table S2. Evaluation criteria for the individual cNNs

Table S3. Comparison of vanilla and bagging ensembles

Table S4. CIN ratios in comparison to *in situ* and cNN results for 22 patients from the UKC cohort

Code S1. Python script