Deep learning trained on lymph node status predicts outcome from gastric cancer histopathology: a retrospective multicentric study

Hannah Sophie Muti, Christoph Röcken, Hans-Michael Behrens, Chiara Maria Lavinia Loeffler, Nic Gabriel Reitsam, Bianca Grosser, Bruno Märkl, Daniel E. Stange, Xiaofeng Jiang, Gregory Patrick Veldhuizen, Daniel Truhn, Matthias P. Ebert, Heike Irmgard Grabsch, Jakob Nikolas Kather

Angaben zur Veröffentlichung / Publication details:

Muti, Hannah Sophie, Christoph Röcken, Hans-Michael Behrens, Chiara Maria Lavinia Loeffler, Nic Gabriel Reitsam, Bianca Grosser, Bruno Märkl, et al. 2023. "Deep learning trained on lymph node status predicts outcome from gastric cancer histopathology: a retrospective multicentric study." *European Journal of Cancer* 194 (September): 113335. https://doi.org/10.1016/j.ejca.2023.113335.

European Journal of Cancer 194 (2023) 113335

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research Article

Deep learning trained on lymph node status predicts outcome from gastric cancer histopathology: a retrospective multicentric study

Hannah S. Muti [a,b], Christoph Röcken [c], Hans-Michael Behrens [c], Chiara M.L. Löffler [a,d], Nic G. Reitsam [e], Bianca Grosser [e], Bruno Märkl [e], Daniel E. Stange [b], Xiaofeng Jiang [a], Gregory P. Veldhuizen [a], Daniel Truhn [f], Matthias P. Ebert [g,h,i,j], Heike I. Grabsch [k,l], Jakob N. Kather [a,d,k,m,*,1]

[a] *Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany*
[b] *Department of Visceral, Thoracic and Vascular Surgery, University Hospital Carl Gustav Carus Dresden, Dresden, Germany*
[c] *Department of Pathology, University Hospital Schleswig-Holstein, Kiel, Germany*
[d] *Department of Medicine I, University Hospital Dresden, Dresden, Germany*
[e] *Pathology, Faculty of Medicine, University of Augsburg, Augsburg, Germany*
[f] *Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Germany*
[g] *Department of Medicine II, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany*
[h] *DKFZ-Hector Cancer Institute at the University Medical Center, Mannheim, Germany*
[i] *Clinical Cooperation Unit Healthy Metabolism, Center for Preventive Medicine and Digital Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany*
[j] *Mannheim Institute for Innate Immunoscience (MI3), Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany*
[k] *Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK*
[l] *Department of Pathology, GROW School for Oncology and Reproduction, Maastricht University Medical Center+, Maastricht, the Netherlands*
[m] *Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany*

Received 29 March 2023; Received in revised form 15 August 2023; Accepted 3 September 2023
Available online 12 September 2023

KEYWORDS
Gastric cancer;
Digital pathology;
Deep learning;
Precision oncology;

Abstract ***Aim:*** Gastric cancer (GC) is a tumour entity with highly variant outcomes. Lymph node metastasis is a prognostically adverse biomarker. We hypothesised that GC primary tissue contains information that is predictive of lymph node status and patient prognosis and that this information can be extracted using deep learning (DL).

* Corresponding author: Fiedlerstraße 74, 01307, Dresden, Germany.
E-mail address: jakob-nikolas.kather@alumni.dkfz.de (J.N. Kather).
[1] +49 351 458 7558

https://doi.org/10.1016/j.ejca.2023.113335

Artificial intelligence

Methods: Using three patient cohorts comprising 1146 patients, we trained and validated a DL system to predict lymph node status directly from haematoxylin and eosin–stained GC tissue sections. We investigated the concordance between the DL-based prediction from the primary tumour slides (aiN score) and the histopathological lymph node status (pN). Furthermore, we assessed the prognostic value of the aiN score alone and when combined with the pN status.

Results: The aiN score predicted the pN status reaching area under the receiver operating characteristic curves of 0.71 in the training cohort and 0.69 and 0.65 in the two test cohorts. In a multivariate Cox analysis, the aiN score was an independent predictor of patient survival with hazard ratios of 1.5 in the training cohort and of 1.3 and 2.2 in the two test cohorts. A combination of the aiN score and the pN status prognostically stratified patients by survival with p-values < 0.05 in logrank tests.

Conclusion: GC primary tumour tissue contains additional prognostic information that is accessible using the aiN score. In combination with the pN status, this can be used for personalised management of GC patients after prospective validation.

1. Introduction

Gastric cancer (GC) has a high incidence and an even higher cancer-specific mortality with a 5-year survival of less than 50% despite optimal multimodal treatment [1]. Whilst early-stage GC is endoscopically resectable, once the tumour has invaded past the submucosal barrier, surgical resection remains the only curative therapy option [2]. Resection protocols include lymphadenectomy of at least 15 lymph nodes, which are subsequently examined by a pathologist to determine the lymph node status (pN status) [3]. The pN status is of prognostic value: patients with lymph node metastases (LNM) show worse survival than patients without LNM, especially in early-stage GC [4]. Conversely, patients with the same pN status can have very different outcomes, which might in part be due to imperfect assessment [5]. The final pN status can depend on the number and sizes of resected lymph nodes; the extent of positively classified lymph nodes can depend on the methods of assessment or fixation or on micrometastases [6–9]. Some histomorphological features in the primary tumour tissue might be indicative of LNM, among which lymphovascular invasion, Laurén subtype or tumour budding; however, this is not taken into consideration for pN status assessment in current diagnostic routines [10,11].

In the past years, deep learning (DL), a method from the field of artificial intelligence (AI), has been used to extract clinically relevant information directly from routine histopathology slides of solid tumours [12,13]. One advantage of AI-based methods is their ability to extract information from haematoxylin and eosin (H&E)-stained tissue slides, a routinely used staining modality [14]. AI-based analysis of H&E slides has been used to diagnose cancer [15], find prognostic information [16,17], predict molecular subtypes [18] or genetic alterations [19] and predict treatment response [20]. Few proof-of-concept studies have investigated the use of AI to predict lymph node status directly from primary tumour histology in colorectal cancer (CRC) [21,22] or from immunohistochemically stained tissue in GC [23]. Other research groups developed AI-based prognostic tools for CRC, of which some received regulatory approval in the European Union and the United States [16,24]. However, very few of these efforts were aimed at GC, despite the high clinical need for better patient stratification in this tumour entity [14].

We hypothesised that DL can detect morphological patterns predictive of the pN status in primary tumour tissue and that these patterns are related to GC patient prognosis (Fig. 1A). Hence, we trained and validated a DL-based classifier to predict the pN status directly from H&E whole slide images (WSIs) of primary GC resection tissue sections. We subsequently investigated the prognostic value of our classifier and analysed the histomorphological features associated with a high prediction score. Finally, we compared the prognostic power of lymph node positivity prediction with the prognostic power of direct survival prediction and proposed an implementation of our biomarker into clinical routine workflows.

2. Materials and methods

2.1. Ethics statement

All experiments were conducted according to the Declaration of Helsinki. This study complies with the 'Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis' statement (Suppl. Table 1) [25]. This study was approved by the ethics boards at RWTH Aachen University Hospital (EK 345/19) and the Medical Faculty of Technical University Dresden (BO-EK-444102022). The collection of patient samples in each cohort was approved at each institution as described below.

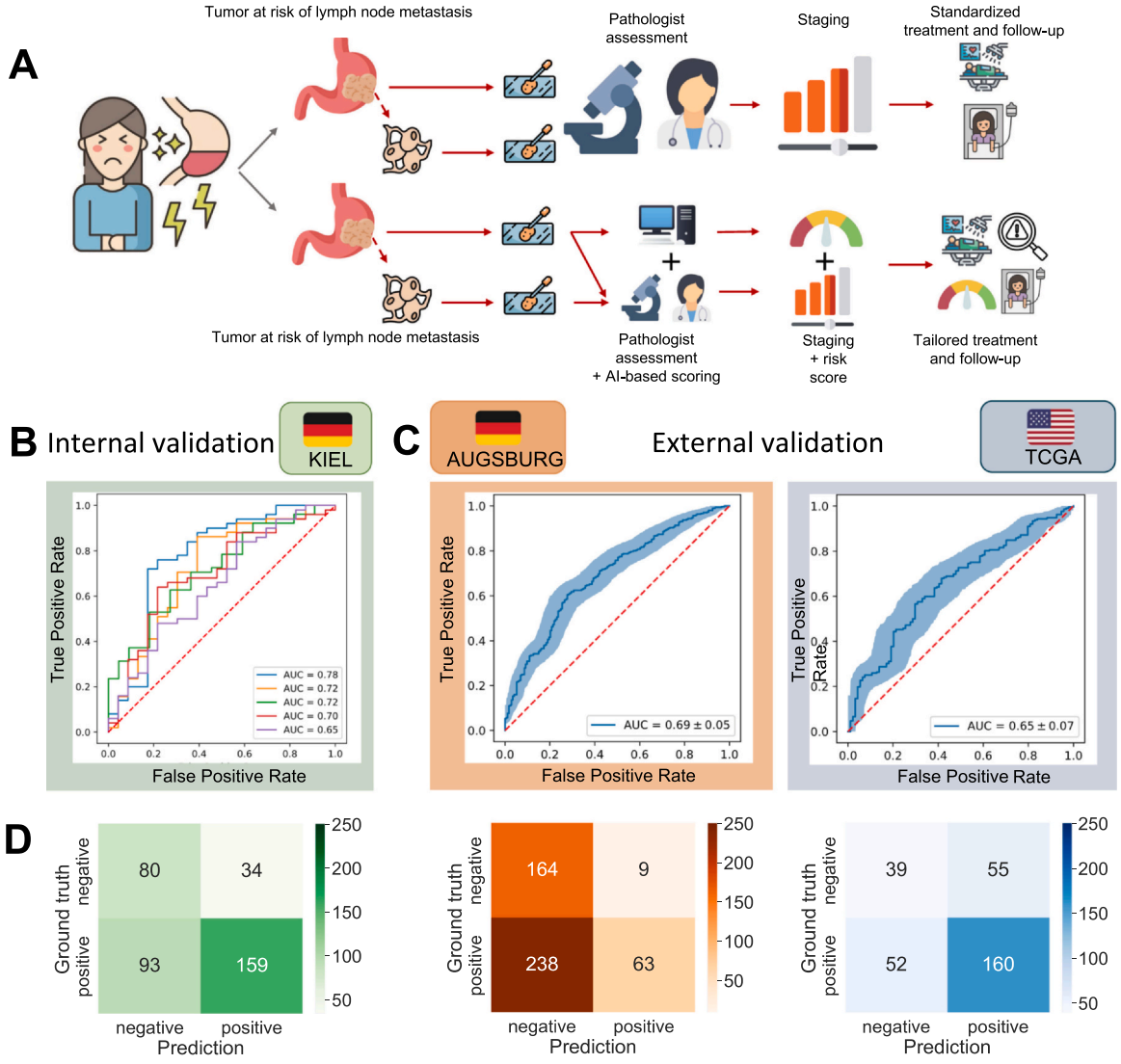

Fig. 1. Hypothesis and key results of this study. (A) Upper panel: current workup of gastric cancer patients. A patient with resectable stomach cancer undergoes surgery, which includes lymphadenectomy. A pathologist analyzes lymph node tissue to determine the lymph node status. On the basis of this, treatment and follow-up decisions are made according to standardised guidelines. Lower panel: AI-augmented workup of gastric cancer patients. A patient with resectable stomach cancer undergoes surgery, which includes lymphadenectomy. A pathologist analyses the lymph node tissue to determine the pN status. Parallely, the primary tissue is analysed by using AI, determining a risk score. Based on the disease stage combined with the risk score, patient management could be personalised. (B) Receiver operating curves (ROC) for the KIEL, AUGSBURG and TCGA cohort. (C) Distribution of model predictions visualised in confusion matrices. AI, artificial intelligence. TCGA, The Cancer Genome Atlas. Pictograms were obtained from https://www.flaticon.com.

2.2. Patient cohorts and data acquisition

We retrospectively collected anonymised, digitised H&E-stained WSI of GC surgical resections from three independent sources. The first cohort originates from the Institute of Pathology at the University Hospital Schleswig Holstein in Kiel, Germany, subsequently referred to as the KIEL cohort ($n = 366$) [26]. The second cohort originates from the Institute of Pathology, University Hospital Augsburg in Augsburg, Germany, subsequently referred to as the AUGSBURG cohort ($n = 474$) [27]. The third cohort is derived from the Cancer Genome Atlas Project, publicly available at https://portal.gdc.cancer.gov and www.cbioportal.org (accessed 10th March 2022) and subsequently referred to as the TCGA cohort ($n = 306$). Only patients for whom both a WSI and pN status (positive/negative) were available were included in our experiments. Throughout all experiments, we used one WSI per patient; in the non-public cohorts, the WSI were pre-selected by board-specified pathologists from the source institutions (C.R., H.M.B., and B.M.). Clinicopathological characteristics for the cohorts are summarised in Table 1.

2.3. Experimental design

We hypothesised that GC primary tumour tissue contains information indicative of the pN status. To detect this, we trained and validated an AI-based classifier

Table 1
Cohort characteristics.

Characteristic	KIEL ($n = 366$)			AUGSBURG ($n = 474$)			TCGA ($n = 306$)		
	n (%)	HR	*p*	*n* (%)	HR	*p*	*n* (%)	HR	*p*
pT									
1	44 (12)	1.0 (ref)	-	69 (15)	1.0 (ref)	-	12 (4)	NA[a]	-
2	33 (9)	1.9	0.04	82 (17)	1.6	0.15	59 (18)	NA[a]	-
3	137 (37)	3.1	< 0.0001	228 (48)	3.0	< 0.0001	135 (44)	NA[a]	-
4	106 (29)	4.9	< 0.0001	95 (20)	6.5	< 0.0001	90 (29)	NA[a]	-
NA	46 (13)	NA	-	0 (0)	NA	-	10 (3)	NA	-
pN									
0	98 (27)	1.0 (ref)	-	173 (37)	1.0 (ref)	-	89 (29)	1.0 (ref)	-
1	43 (11)	2.1	0.0007	83 (17)	1.5	0.05	82 (27)	1.7	0.07
2	56 (15)	2.7	< 0.0001	96 (20)	2.3	< 0.0001	62 (20)	1.7	0.08
3	123 (34)	4.2	< 0.0001	122 (26)	3.5	< 0.0001	63 (21)	3.2	< 0.0001
NA	46 (13)	NA	-	1 (0)	NA	-	10 (3)	NA	-
pM									
0	259 (71)	1.0	-	344 (73)	1.0 (ref)	-	269 (88)	1.0 (ref)	-
1	61 (16)	3.0	< 0.0001	98 (20)	0.9	0.69	17 (9)	3.4	< 0.0001
NA	46 (13)	NA	-	32 (7)	NA	-	10 (3)	NA	-
Sex									
M	210 (57)	1.1	0.32	313 (66)	1.0	0.76	196 (64)	1.4	0.12
F	110 (30)	1.0 (ref)	-	161 (34)	1.0 (ref)	-	100 (33)	1.0 (ref)	-
NA	46 (13)	NA	-	1 (0)	NA	-	10 (3)	NA	-
Laurén grade									
Intestinal	187 (51)	0.8	0.05	276 (58)	0.8	0.23	129 (42)	0.5	0.3
Diffuse	75 (21)	1.0	-	145 (31)	1.0	-	52 (17)	1.0 (ref)	-
Mixed	20 (6)	1.1	0.81	53 (11)	1.3	0.20	14 (5)	1.9	0.10
Other/NA	84 (22)	NA	-	0 (0)	NA	-	111 (36)	NA	-
UICC stage									
I	55 (15)	1.0	-	86 (18)	1.0 (ref)	-	38 (12)	1.0 (ref)	-
II	72 (19)	2.0	0.005	116 (25)	1.9	0.005	92 (30)	1.8	0.20
III	132 (36)	4.3	< 0.0001	176 (37)	3.8	< 0.0001	137 (45)	2.8	0.01
IV	61 (17)	8.1	< 0.0001	96 (20)	5.5	< 0.0001	27 (9)	5.8	< 0.0001
NA	46 (13)	NA	-	0 (0)	NA	-	12 (4)	NA	-
Neoadjuvant treatment									
Yes	0 (0)	NA	-	135 (28)	0.8	0.06	0 (0)	NA	-
No	366 (100)	NA	-	339 (72)	1.0 (ref)	-	306 (100)	NA	-
Tumour budding									
Bd0	90 (25)	1.0 (ref)	-	0 (0)	NA	-	0 (0)	NA	-
Bd1	61 (16)	2.0	0.001	0 (0)	NA	-	0 (0)	NA	-
Bd2	24 (6)	2.5	0.0007	0 (0)	NA	-	0 (0)	NA	-
Bd3	151 (42)	2.3	< 0.0001	0 (0)	NA	-	0 (0)	NA	-
NA	40 (11)	NA	-	474 (100)	NA	-	306 (100)	NA	-
SARIFA									
Positive	0 (0)	NA	-	96 (20)	2.0	< 0.0001	0 (0)	NA	-
Negative	0 (0)	NA	-	378 (80)	1.0 (ref)	-	0 (0)	NA	-
NA	366 (100)	NA	-	0 (0)	NA	-	306 (100)	NA	-
Survival									
OS in months, median (IQR)	15.52 (7.20, 36.33)			19.00 (6.00, 46.75)			16.37 (9.58, 27.40)		
Event rate in 5 years	0.67			0.49			0.37		
Censoring rate in 5 years	0.18			0.33			0.57		
Scanner type (resolution)	Hamamatsu (0.22 µm/pixel)			3DHistech (0.27 µm/pixel)			Leica Aperio (0.25 µm/pixel)		

F, female; HR, hazard ratio; IQR, interquartile range; M, male; *n, number of cases included in our experiments; NA, not available; OS, overall survival; ref, reference; SARIFA, Stroma AReactive Invasion Front Areas; TCGA, The Cancer Genome Atlas; UICC, Union Internationale Contre le Cancer.*

All statistical analyses were performed based on the available data as depicted in this table, excluding NA cases.

[a] Not available: model did not converge due to the absence of deaths in the reference group (T1).

(aiN score) to predict the pN status from WSI of GC H &E-stained resection specimens. First, we performed an internal validation in a patient-level within-cohort cross-validated design in fivefolds on the KIEL cohort. To ensure the generalisability of our model, we trained a model on KIEL and externally validated it on two independent test sets, the AUGSBURG and TCGA cohorts. During all training and testing, no patient was

ever present in train and test sets simultaneously. Furthermore, we hypothesised that a higher aiN score would be associated with a shorter overall survival (OS). We performed a survival analysis comparing patients who were predicted to have lymph node metastasis (aiN +) and patients who were predicted not to have lymph node metastasis (aiN−) and in risk groups created from a combination of the aiN and the pN status. Third, we analysed the histomorphology of aiN+ cases to identify histopathological characteristics that were indicative of lymph node metastasis to our model. Finally, we compared the prognostic value of our lymph node status classifier with a classifier that was trained to predict OS after 5 years from the date of surgery.

2.4. DL procedures

We used our open-source end-to-end DL pipeline to train a neural network to predict the pN status from primary GC WSI. As a preprocessing step, we tessellated the WSI into tiles of 256×256 μm, which we processed at 224 px edge length, yielding a resolution of 1.14 μm per pixel. We colour normalised the tiles using the Macenko method and discarded blurry tiles using the Canny edge detection as described before [28,29]. We extracted tile-wise feature vectors with RetCCL, a neural network pretrained on a histopathology data set [30]. RetCCL is based on clustering-guided contrastive learning with the objective of maximising the contrastive loss between tiles that originate from different images in a self-supervised manner. The resulting feature vectors function as the input for an attention-based multiple instance learning network (attMIL), which was our final classification network, as described before [31–34]. In attMIL, the feature vectors are grouped on a patient-level with a patient-level label, which the model is finally trained on. During training, the impact of a single feature vector on the final classification is calculated, which we refer to as the attention. The attention, through which the model learns the importance of certain features for the classification task, is included in the aggregation step. This has the effect that areas of high attention weigh more into the final prediction than areas of low attention. The classification problem was binary (positive or negative), and the ground truth was obtained from the original pathology report. The attMIL model has two output neurons, in our case N+ and N−, the activations of which are then softmaxed to sum up to 1. The activation of the 'N+' output neuron was used as the 'aiN score' when deploying the trained model on unseen WSIs. For subsequent statistical tests, we binarised the continuous model prediction at a prespecified cutoff of 0.5 to obtain an unbiased estimate of external validation performance. To visualise our model output, we created patient-wise high-resolution heatmaps from the feature vectors extracted with RetCCL, displaying the region-wise attention and region-wise prediction scores. The heatmaps were then blended with the original WSI, with red areas indicating high attention/aiN+ and blue indicating low attention/aiN−.

2.5. Statistical analysis

The model performance was primarily assessed by the area under the receiver operating characteristic curves (AUROCs) with 95% confidence intervals (CIs). In the internal validation, the CI was obtained through all fivefold-wise AUROCs, and in the external validation, CIs were obtained through 1000-fold bootstrapping. The statistical power of the AUROCs was assessed through two-sided unpaired t-tests comparing the prediction scores, with the null hypothesis being that the aiN scores for the positive and negative groups (according to the ground truth) are from the same distribution. The relationship between the pN status, the aiN score and a combination thereof with the OS was assessed using Kaplan-Meier analyses with logrank statistics, Harrel's concordance indices (c-indices), and univariate and multivariate Cox proportional hazard models for OS. In the multivariate Cox proportional hazard models, age, sex and staging parameters, including 'Union Internationale Contre le Cancer' (UICC) and T-, N- and M-stages, were included as covariates alongside the 'aiN score'. The association between our aiN score and morphological characteristics was analysed using Kruskal-Wallis and Dunn's tests. Cases that were included in our model training and testing but lacked other clinicopathological data were excluded from statistical analysis. For statistical calculations, we used the R 'survminer' and 'intsurv' libraries. p-values ≤0.05 were considered significant, and no correction for multiple testing was applied.

2.6. Code availability

All source codes for image preprocessing are publicly available at https://github.com/KatherLab/preProcessing. All source codes for training and evaluating DL models are publicly available at https://github.com/KatherLab/marugoto. All source codes for the high-resolution heatmaps are available at https://github.com/KatherLab/highres-WSI-heatmaps.

3. Results

3.1. DL only moderately predicts the pN status from primary tumour resection slides in GC

We trained and validated a classifier to predict the pN status from H&E-stained WSI of primary GC resection tissue specimens in our three patient cohorts with attMIL. In the internal validation, we achieved an AUROC of 0.71 (95% CI ± 0.06, $p = 0.01$) within KIEL (Fig. 1B) In the external validation cohorts, we achieved

Table 2
AUROC and Cox proportional hazard analyses.

Cohort	KIEL			AUGSBURG			TCGA		
	AUROC	CI	*p*	AUROC	CI	*p*	AUROC	CI	*p*
AUROC analysis	0.7145	± 0.06	0.01	0.6924	± 0.05	< 0.0001	0.657	± 0.07	0.0001
C-index (pN)	0.61			0.59			0.58		
C-index (aiN)	0.61			0.54			0.56		
Covariates[a]	**HR**	**CI**	***p***	**HR**	**CI**	***p***	**HR**	**CI**	***p***
Prediction	2.3	1.8;3.0	< 0.0001	1.9	1.4;2.5	0.0002	2.3	1.4;3.6	0.0007
Age	1.0	1.0;1.0	0.15	1.0	1.0;1.1	< 0.0001	1.0	1.0;1.1	0.0004
UICC	1.9	1.6;2.2	< 0.0001	1.9	1.7;2.2	< 0.0001	1.8	1.4;2.4	< 0.0001
Sex	1.2	0.9;1.6	0.21	0.8	0.6;1.1	0.2	1.3	0.8;1.9	0.27
Prediction	1.5	1.1;1.9	0.008	1.3	1.0;1.9	0.08	2.2	1.4;3.6	0.001
Age	1.0	1.0;1.0	0.29	1.0	1.0;1.0	< 0.0001	1.0	1.0;1.1	0.003
T stage	1.5	1.3;1.8	< 0.0001	1.9	1.6;2.3	< 0.0001	1.5	1.2;2.0	0.0008
Sex	1.3	1.0;1.7	0.08	1.2	0.9;1.6	0.18	1.4	0.9;2.1	0.10
Prediction	1.6	1.2;2.2	0.0005	1.4	1.0;1.9	0.04	2.3	1.4;3.7	0.0007
Age	1.0	1.0;1.0	0.40	1.0	1.0;1.0	< 0.0001	1.0	1.0;1.1	0.004
N Stage	1.5	1.3;1.7	< 0.0001	1.5	1.4;1.7	< 0.0001	1.4	1.2;1.6	0.0002
Sex	1.1	0.9;1.5	0.36	1.1	0.9;1.5	0.41	1.3	0.9;2.0	0.20
Prediction	1.7	1.3;2.3	< 0.0001	1.2	0.9;1.7	0.233	2.2	1.4;3.6	0.001
Age	1.0	1.0;1.0	0.34	1.0	1.0;1.0	< 0.0001	1.0	1.0;1.1	0.002
M stage	2.4	1.8;3.3	< 0.0001	0.9	0.7;1.3	0.69	4.3	2.3;8.2	< 0.0001
Sex	1.2	0.9;1.6	0.15	1.0	0.8;1.4	0.77	1.5	1.0;2.3	0.002
Prediction	2.0	1.5;2.6	< 0.0001	1.8	1.3;2.4	0.0003	2.3	1.4;3.7	0.0007
Age	NA	NA	NA	1.0	1.0;1.0	< 0.0001	NA	NA	NA
Pre-treatment	NA	NA	NA	0.9	0.7;1.2	0.4	NA	NA	NA
Sex	NA	NA	NA	1.0	0.8;	0.7	NA	NA	NA
Prediction	NA	NA	NA	1.8	1.3;2.4	0.0003	NA	NA	NA

aiN, aiN score; AUROC, area under the receiver operating characteristic curve; CI, confidence interval; HR, hazard ratio; NA, not available; pN, lymph node status (positive versus negative) as assessed by pathologist; TCGA, The Cancer Genome Atlas; UICC, Union Internationale Contre le Cancer.
[a] The sample size of the sets in the case of the multivariate Cox analyses is equal to the number of cases for which the respective covariate was available.

AUROCs of 0.69 (95% CI ± 0.05, $p < 0.0001$) in AUGSBURG and 0.65 (95% CI ± 0.07, $p < 0.0001$) in TCGA (Fig. 1C) Confusion matrices for the distribution of results at a prespecified cutoff of 0.5 are shown in Fig. 1D. All results are summarised in Table 2. We concluded that our classifier only moderately predicts the pN status.

3.2. The lymph node–status–derived aiN score is an independent predictor of OS in GC

As the pN status is related to patient survival, we hypothesised that the patterns identified as indicative of LNM from primary GC tissue by our model would be prognostically relevant for GC patients. Therefore, we investigated the prognostic values of the pN status and our aiN score by performing Kaplan-Meier analyses with logrank tests to compare the survival stratification. The pN status (positive versus negative) stratified patients in all cohorts ($p < 0.0001$ in KIEL and AUGSBURG; $p = 0.002$ in TCGA; Fig. 2A). Similarly, the aiN score was able to stratify patients by survival in all three cohorts (all *p*-values < 0.0001; Fig. 2B, Table 2) We subsequently calculated the c-indices and found that the pN status and the aiN score yield similar c-indices of 0.61, 0.59 and 0.58 according to the pN status and of 0.61, 0.54 and 0.56 according to the aiN score in KIEL, AUGSBURG and TCGA (Table 2). Furthermore, we conducted Cox proportional hazard analyses with age, sex and UICC stage of a given patient as covariates. In KIEL, the aiN score yielded a univariate HR of 2.28 (95% CI ± 0.13, $p < 0.0001$; Table 2) and a multivariate HR of 1.5 (1.1, 1.9; $p = 0.0080$; Fig. 2C). In AUGSBURG, the aiN score yielded a univariate HR of 1.86 (95% CI ± 0.6, $p = 0.0050$, Table 2) and a multivariate HR of 1.3 (1.0;1.9, $p = 0.05$, Fig. 2C). In TCGA, the aiN score yielded a univariate HR of 2.25 (95% CI ± 0.24, $p = 0.0007$, Table 2) and a multivariate HR of 2.2 (1.4;3.6, $p = 0.0012$, Fig. 2C). We repeated the multivariate Cox proportional hazard analyses with the individual staging parameters pT-, pN- and M-stage and the pretreatment status as covariates in all eligible cohorts, where our score was evenly shown to be an independent prognostic predictor (Table 2). To conclude, the aiN score proved to predict survival independent of known prognostic factors. Together, our results show that although our aiN score was not perfectly concordant with the pN status, it is predictive of survival.

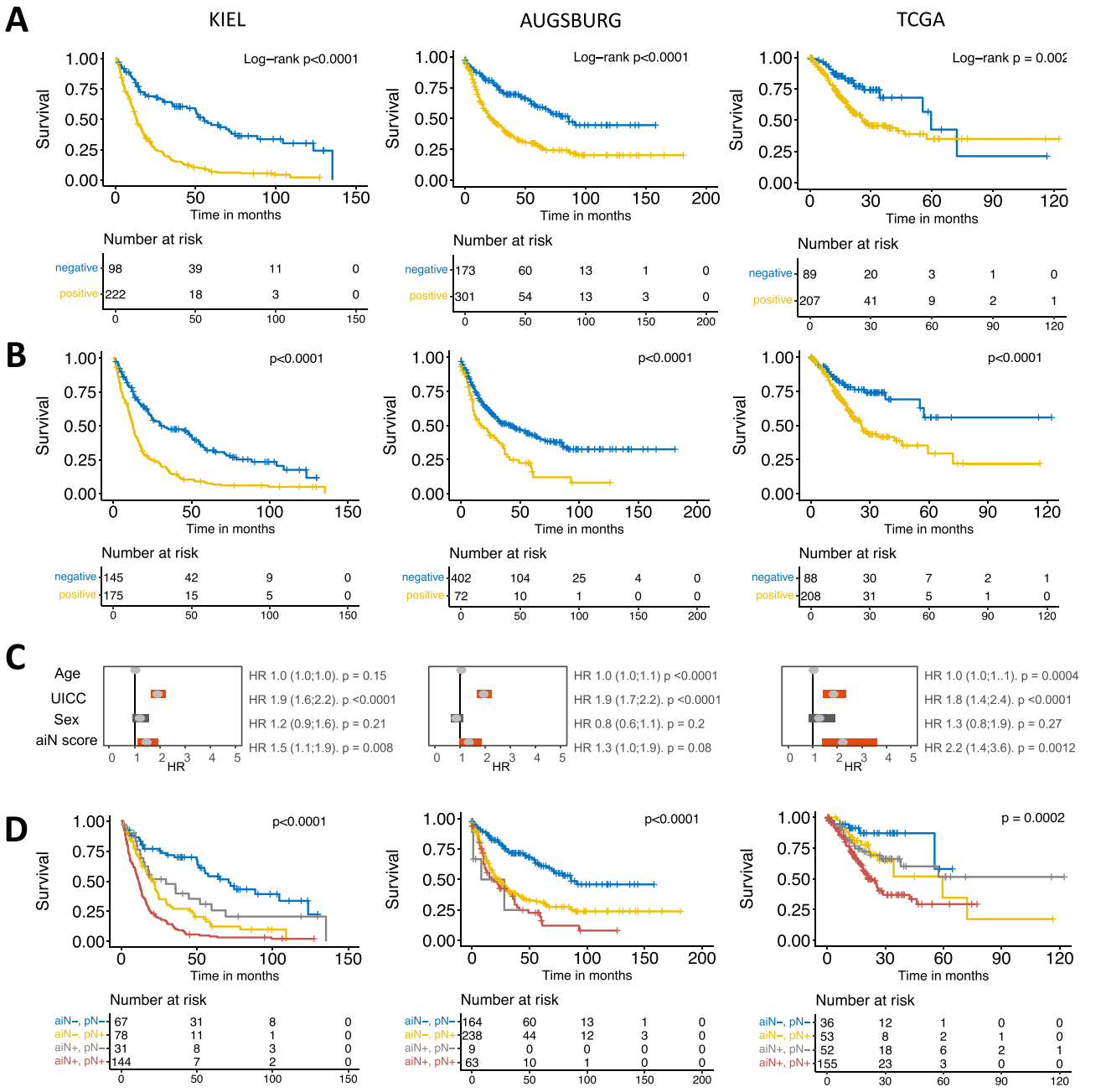

Fig. 2. Artificial intelligence–generated lymph node metastasis score augments patient stratification. (A) Kaplan-Meier plots showing survival stratification by positive and negative pN status in pairwise logrank tests. (B) Kaplan-Meier plots showing survival stratification by positive and negative aiN score in pairwise logrank tests. (C) Forest plots obtained from multivariate Cox proportional hazard analyses with age, sex and UICC disease stage as covariates. (D) Kaplan-Meier plots showing survival stratification by the combination of the aiN score and pN status in combined logrank tests. aiN−, pN−, negative by our model, negative by pathologist assessment; aiN−, pN+, negative by our model, positive by pathologist assessment; aiN+, pN−, positive by our model, negative by pathologist assessment; aiN+, pN+, positive by our model, positive by pathologist assessment; HR, hazard ratio; OS, overall survival; TCGA, The Cancer Genome Atlas; UICC, Union Internationale Contre le Cancer.

3.3. The aiN score can augment prognostic stratification of GC patients

Based on the previous results, we investigated whether the aiN score can further stratify patients into high and low risk subgroups that would be of prognostic value. We analysed patient survival by combining our score (aiN+ or aiN−) with the pathologist assessment (pN+ or pN−). We observed differences in the OS when patients were stratified by this combined classification: aiN−, pN− significantly patients showed the best and aiN+, pN+ the worst survival in all our cohorts (Fig. 2D, Suppl. Table 2). In KIEL and TCGA, the aiN+, pN− patients had poorer survival than aiN−,

pN− patients (Fig. 2D). In AUGSBURG, the survival of the aiN+, pN− patients was even equal to the survival of the aiN+, pN+ group (Fig. 2D). To conclude, there might be a subset of patients at risk of poor outcome, which cannot be identified using current disease staging methods. Combining existing classification systems, such as the Tumor - Node - Metastasis (TNM) and UICC staging system, with the aiN score could enhance the stratification of GC patients.

3.4. The morphological features related to lymph node metastasis and poor outcome in GC primary tumour resection specimen according to the aiN score

We subsequently assessed the morphology highlighted as indicative of LNM by our model. We utilised high-resolution heatmaps to visually display the spatial distribution of attention and classification scores of the model across a WSI in our validation cohorts. High-attention areas included mostly tumour tissue, whereas other tissue types were less relevant for the aiN score (Fig. 3A and B). Areas of desmoplastic stromal reaction, tumour budding and fat cells adjacent to tumour cells without desmoplastic reaction, described as Stroma AReactive Invasion Front Areas (SARIFA) in previous literature [27], were highlighted as characteristic of aiN+ tumours (Fig. 3A and B).

We consecutively explored the correlation between the Laurén grade, the presence of tumour budding, SARIFA and our aiN score. In KIEL and AUGSBURG, we examined whether the aiN score differed between diffuse and non-diffuse GC cases. With p-values of $p = 0.06$ for the KIEL cohort and $p = 0.41$ for the AUGSBURG cohort, there was no significant difference in the aiN score between the two groups (Fig. 3C and D). Tumour budding information was available in the KIEL cohort, classified as no buds (Bd0) to > 10 buds (Bd3) [11]. Tumour budding was significantly associated with a higher aiN score ($p < 0.0001$), and Bd1–3 patients had a significantly higher aiN score than Bd0 patients ($p = 0.014$, $p = 0.0005$ and $p < 0.0001$, Fig. 3E). SARIFA status (positive versus negative) was available in the AUGSBURG cohort. A positive SARIFA status was significantly associated with a higher aiN score ($p = 0.0008$, Fig. 3F). To conclude, these data suggest that our classifier potentially identifies morphological characteristics in the primary tumour that are indicative of LNMs.

3.5. The aiN score in comparison with prognostic strata

Next, we examined whether the survival stratification through the aiN score was dependent on the pT stage. In KIEL and AUGSBURG, the survival differences were significant in pT1 and pT2 patients ($p = 0.002$ and $p < 0.0001$, Fig. 4A and B). In TCGA, the survival differences were significant in pT3 and pT4 patients ($p = 0.0003$, Fig. 4C). A combination of the aiN score and pN status was predictive of survival in the pT1&2 subset of the KIEL and AUGSBURG cohorts ($p = 0.0011$ and $p = 0.0001$, Fig. 4D and E) and in the T3 &4 subset of the TCGA cohort ($p = 0.0003$, Fig. 4F). We conclude that our model output is not dependent on the pT stage. In addition, the number of aiN-pN+ patients was higher in AUGSBURG compared to other cohorts (Fig. 1D). Since a subset of the AUGSBURG cohort had received neoadjuvant chemotherapy, we compared our prediction with the pretreatment status and found that the percentage of aiN-pN+ cases is higher in the pretreated group. Finally, we compared the prognostic value of our aiN score with that of direct survival prediction. We binarised the OS after 5 years into dead or alive for each patient and trained a survival classifier. Using the binarised survival prediction score, we performed multivariate Cox proportional hazard analyses for the OS outcome categories and found that only in TCGA, the survival classifier was an independent survival predictor with a multivariate HR of 1.8 (1.2; 2.5, $p = 0.002$; Suppl. Fig. 1). Comparing this to our aiN classification results, the aiN score yields a robust prognostic indication from GC primary resection specimen. A DL-based prognostication system, such as the aiN score, could in the future be integrated into GC patient management in the form of a risk score. Combining the aiN score and the pN status, GC patients can be stratified into high, intermediate and low risk groups, and personalised workup or management strategies can be applied (Fig. 4H).

4. Discussion

In the past decades, numerous clinical trials attempted to innovate on therapeutic strategies for GC. Apart from PD-L1 and *HER2*-positive patients, surgery and chemotherapy remain the only treatment options for the majority of GC patients [36]. Despite the broad variety of clinical outcomes in GC patients, tumour staging remains the indicator for therapy allocation [2]. Hence, new biomarkers are needed to enable a more targeted delivery of different treatment modalities.

Here, we developed and validated the aiN score, a DL-based biomarker derived from the pN status trained on H&E-stained WSI of the primary tumour for prognostication in GC. Based on the prediction of AUROCs, the aiN score cannot be used to replace pN status assessment as performed by pathologists but predicted survival independently from clinicopathological staging parameters.

Adding to the pN status, we were able to stratify GC patients into subgroups with high risk (aiN+) and low risk (aiN−) of poor survival solely based on primary tumour tissue. This suggests that primary tumour tissue contains prognostic information that is neglected in the current clinicopathological workup but can be exploited by DL to augment the diagnostic process.

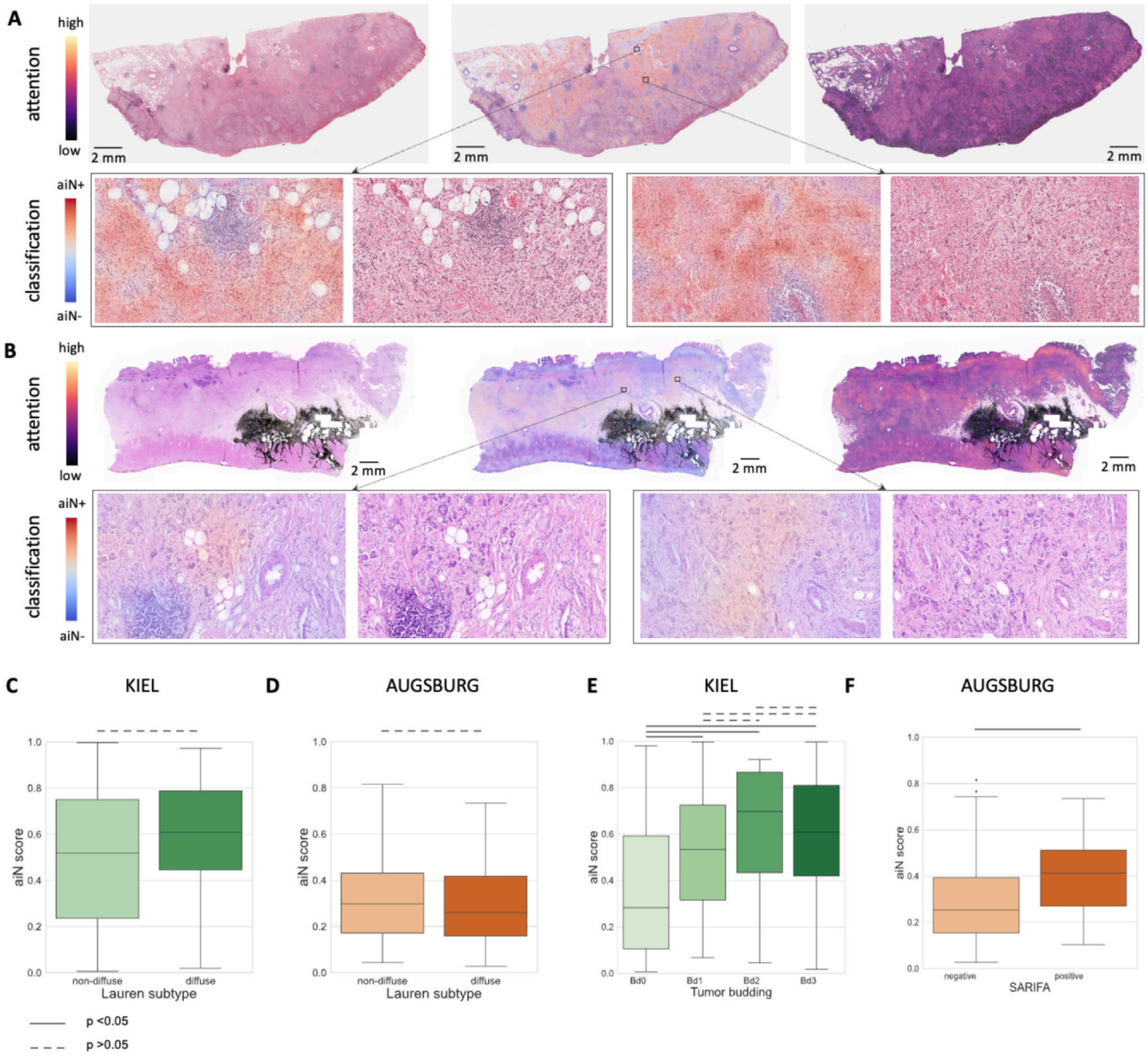

Fig. 3. The association between morphological characteristics and the aiN score. (A) Left to right: H&E-stained slide, classification heatmap and attention heatmap for an aiN+pN+ T3 patient from TCGA. The classification heatmap depicts stromal reaction highlighted as predictive of lymph node metastasis. (B) Left to right: H&E-stained slide, classification heatmap and attention heatmap for an aiN +pN− T2 patient of the AUGSBURG cohort. The classification heatmap depicts stromal reaction and SARIFA highlighted as predictive of lymph node metastasis by our model. (C) Boxplots showing aiN scores according to the Laurén grade in the KIEL cohort. (D) Boxplots showing aiN scores according to the Laurén grade in the AUGSBURG cohort. (E) Boxplots showing aiN scores according to tumour budding in the KIEL cohort. (F) Boxplots showing aiN scores according to SARIFA status in the AUGSBURG cohort (right). H &E, haematoxylin and eosin; SARIFA, Stroma AReactive Invasion Front Areas; TCGA, The Cancer Genome Atlas.

We analysed the histomorphology of aiN+ cases and found SARIFA and tumour budding, which have been linked to poor outcome in gastrointestinal cancers before, to be significantly associated with a higher aiN score [11,27,37–39]. Tumour budding and clusters of poorly differentiated cells were previously found to be indicative of LNMs [11,40], and the SARIFA status was not associated with a positive pN status in GC [27]. Another feature highlighted by our model was desmoplastic stroma reaction, which has been identified as prognostically adverse in a recent clinical trial in CRC [41]. Whether our model indicates a future potential to metastasize to the lymph nodes or whether the aiN score is actually associated with micrometastatic disease is to be determined; however, our results show that DL-based scoring can indeed contain prognostic information that adds to current clinical risk stratification approaches.

Lately, DL-based prognostication approaches emerged, which were mostly aimed at CRC [16,22]. In GC, few studies have aimed at AI-assisted prognostication so far. Wang et al. developed an approach to

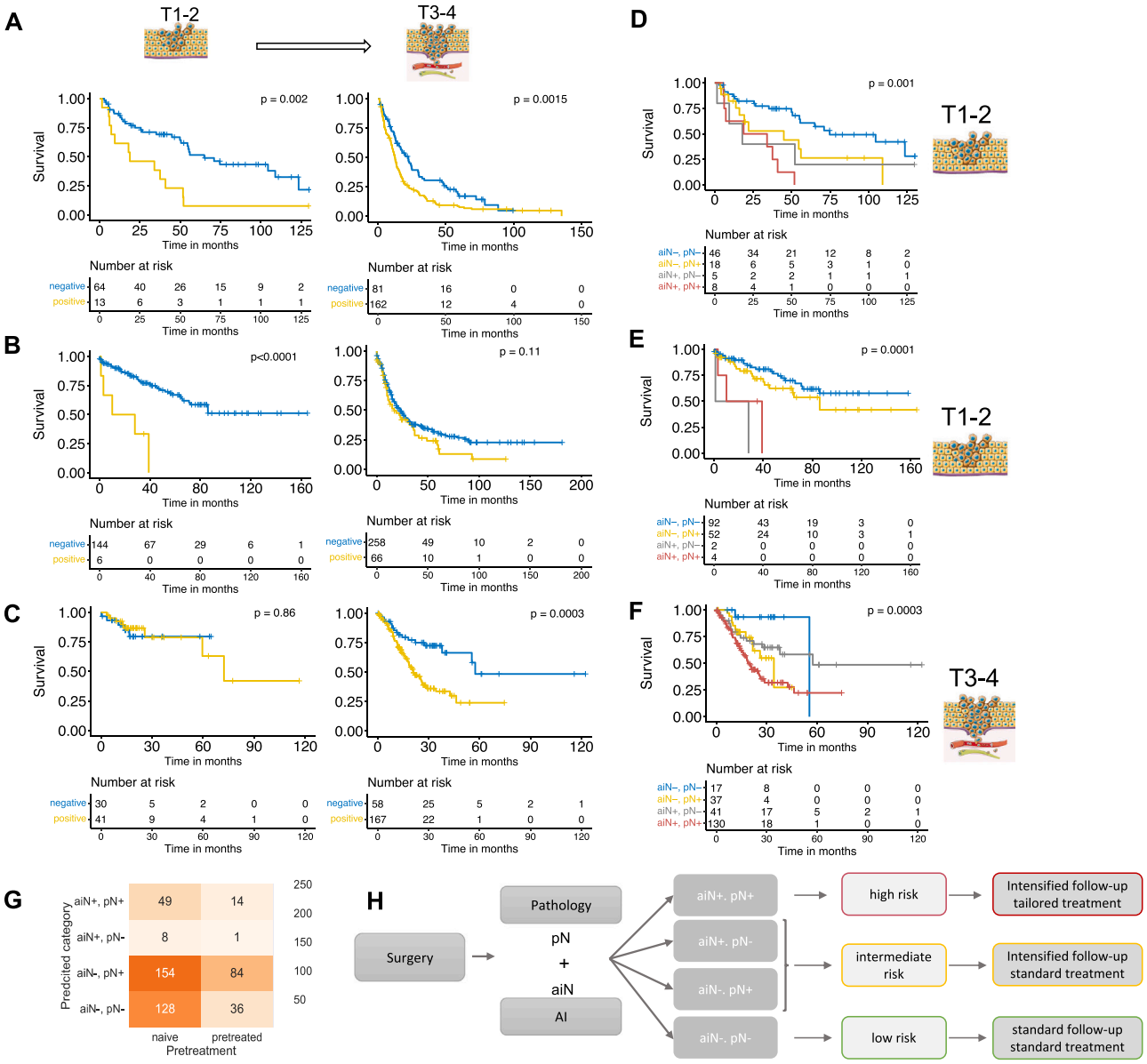

Fig. 4. The prognostic value of the aiN score according to T category and envisioned clinical implementation. Kaplan-Meier plots visualising survival stratification through the aiN score in early and late T stages in (A) the KIEL cohort (internal validation), (B) the AUGSBURG cohort (external validation), (C) TCGA cohort (external validation), (D) Kaplan-Meier plots visualising survival stratification by a combination of the aiN score and the pN status in T1 and T2 patients in the KIEL cohort (internal validation), (E) Kaplan-Meier plots visualising survival stratification by a combination of the aiN score and the pN status in T1 and T2 patients in the AUGSBURG cohort (external validation), (F) Kaplan-Meier plots visualising survival stratification by a combination of the aiN score and the pN status in T3 and T4 patients in the TCGA cohort (external validation), (G) confusion matrix showing the number of patients per prediction category according to the pretreatment status in the AUGSBURG cohort, (H) envisioned workflow implementing conventional staging by pathologists and a DL-based risk scoring system into clinical routine. aiN−, pN−, negative by our model, negative by pathologist assessment; aiN−, pN+, negative by our model, positive by pathologist assessment; aiN+, pN−, positive by our model, negative by pathologist assessment; aiN+, pN+, positive by our model, positive by pathologist assessment. The pictograms were obtained from https://smart.servier.com.

assess the tumour-to-metastasis ratio from resected lymph nodes using DL, which was shown to be an independent predictor of survival [42]. Compared to our experimental setup, they use histomorphology to extract prognostic information without direct clinical implications, whereas our approach connects human and AI-based assessment to prognostically stratify GC patients and connects the histomorphology of the primary to the presence of LNMs. Another group used DL to directly predict survival from GC WSI; however, their model did not prove to be a clear and independent predictor of survival in an external test set [43]. In the end, an

AI-based risk score could conceivably be integrated into GC patient management in the form of a decision aid. Prospective overlap studies using larger cohorts are needed to confirm the prognostic impact and identify vulnerable patient subsets for improving patient survival with AI-based support in GC.

Finally, our aiN score was not an accurate predictor of the pN status according to our prediction AUROCs, yet proved to be an independent predictor of survival in our Cox models. The AUROC is a standard performance metric of AI-based studies in computational pathology but might not accurately reflect the model's ability to generalise if used as the only end-point [35]. In line with our results, model performance assessment should not only be limited to an AUROC analysis, and the use of additional performance metrics in comparable studies should be encouraged.

4.1. Limitations

Our study has several limitations. First, patient-specific factors, such as the performance status, may impact prognosis in addition to lymph node status, which our approach does not reflect. Second, despite our model clearly containing prognostic information, the clinical relevance and therapeutic implications thereof would still have to be evaluated at a larger scale, at best in prospective validation studies. In addition, a well-salvageable subset of GC patients are pT1 patients, who are eligible for endoscopic tumour mucosal resection. Unfortunately, due to a low number of cases, we were not able to assess the prognostic power of our model in this specific subset of patients. Finally, we have compared our aiN score to direct survival prediction from primary tumour tissue using the same method. However, due to censored cases, the number of patients included in the latter is lower, which impairs the interpretability of these results.

4.2. Conclusion

Overall, our study shows that the primary tumour tissue of GC resection specimens contains prognostic information that is neglected in the current diagnostic process but accessible through DL-based analysis as suggested by us. In the future, this information could be implemented in the clinical workup of GC cases and provide a risk stratification of GC patients.

Funding

J.N.K. is supported by the German Federal Ministry of Health (DEEP LIVER, ZMVI1-2520DAT111) and the Max-Eder-Programme of the German Cancer Aid (grant #70113864), the German Federal Ministry of Education and Research (PEARL, 01KD2104C), and the German Academic Exchange Service (SECAI, 57616814).

HIG and JNK are supported in part by the National Institute for Health and Care Research (NIHR) Leeds Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

CRediT authorship contribution statement

H.S.M., J.N.K. and H.I.G. conceptualised the study. C.R., H.M.B., B.M. and B.G. provided the data. J.N.K. provided funding, resources and methodology. H.S.M., B.G. and C.M.L.L. curated the data. H.S.M. conducted the experiments and analysed the data. H.S.M. and C.M.L.L. visualised the results. H.S.M., N.G.R. and H.I.G. revised the heatmaps. H.S.M. wrote the original draft of the manuscript. All authors jointly validated the data, critically revised the manuscript and agreed to the submission of this article.

Declaration of Competing Interest

H.I.G. has received fees for advisory board activities by AstraZeneca and BMS, not related to this study. J.N.K. reports consulting services for Owkin, France, Panakeia, UK, and DoMore Diagnostics, Norway, furthermore J.N.K. holds shares in StratifAI GmbH and has received honoraria for lectures by AstraZeneca, Bayer, Eisai, MSD, BMS, Roche, Pfizer and Fresenius, not related to this study. B.M. has received compensations of travel expenses and fees for advisory board activities by AstraZeneca, Boehringer Ingelheim, MERCK, MSD, BMS, Bayer and Novartis, not related to this study. All other authors declare no conflicts of interest.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ejca.2023.113335.

References

[1] Smyth EC, Wotherspoon A, Peckitt C, Gonzalez D, Hulkki-Wilson S, Eltahir Z, et al. Mismatch repair deficiency, microsatellite instability, and survival: an exploratory analysis of the Medical Research Council Adjuvant Gastric Infusional Chemotherapy (MAGIC) trial. JAMA Oncol 2017;3:1197–203.
[2] Smyth EC, Nilsson M, Grabsch HI, van Grieken NCT, Lordick F. Gastric cancer. Lancet 2020;396:635–48.
[3] Degiuli M, De Manzoni G, Di Leo A, D'Ugo D, Galasso E, Marrelli D, et al. Gastric cancer: current status of lymph node dissection. World J Gastroenterol 2016;22:2875–93.
[4] Gotoda T, Yanagisawa A, Sasako M, Ono H, Nakanishi Y, Shimoda T, et al. Incidence of lymph node metastasis from early gastric cancer: estimation with a large number of cases at two large centers. Gastric Cancer 2000;3:219–25.
[5] Coburn NG. Lymph nodes and gastric cancer. J Surg Oncol 2009;99:199–206.

[6] Wu Z-Y, Li J-H, Zhan W-H, He Y-L, Wan J. Effect of lymph node micrometastases on prognosis of gastric carcinoma. World J Gastroenterol 2007;13:4122–5.
[7] Maehara Y, Oshiro T, Endo K, Baba H, Oda S, Ichiyoshi Y, et al. Clinical significance of occult micrometastasis lymph nodes from patients with early gastric cancer who died of recurrence. Surgery 1996;119:397–402.
[8] Hayashi N, Ito I, Yanagisawa A, Kato Y, Nakamori S, Imaoka S, et al. Genetic diagnosis of lymph-node metastasis in colorectal cancer. Lancet 1995;345:1257–9.
[9] Herrera L, Villarreal JR. Incidence of metastases from rectal adenocarcinoma in small lymph nodes detected by a clearing technique. Dis Colon rectum 1992;35:783–8.
[10] Roviello F, Rossi S, Marrelli D, Pedrazzani C, Corso G, Vindigni C, et al. Number of lymph node metastases and its prognostic significance in early gastric cancer: a multicenter Italian study. J Surg Oncol 2006;94:275–80. discussion 274.
[11] Ulase D, Heckl S, Behrens H-M, Krüger S, Röcken C. Prognostic significance of tumour budding assessed ingastric carcinoma according to the criteria of the International Tumour BuddingConsensus Conference. Histopathology 2020;76:433–46.
[12] Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. Nat Rev Clin Oncol 2019;16:703–15.
[13] Shmatko A, Ghaffari Laleh N, Gerstung M, Kather JN. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. Nat Cancer 2022;3:1026–38.
[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;124:686–96.
[15] Kleppe A, Skrede O-J, De Raedt S, Liestøl K, Kerr DJ, Danielsen HE. Designing deep learning studies in cancer diagnostics. Nat Rev Cancer 2021;21:199–211.
[16] Kleppe A, Skrede O-J, De Raedt S, Hveem TS, Askautrud HA, Jacobsen JE, 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;23:1221–32.
[17] Skrede O-J, De Raedt S, Kleppe A, Hveem TS, Liestøl K, Maddison J, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. Lancet 2020;395:350–60.
[18] Coudray N, Tsirigos A. Deep learning links histology, molecular signatures and prognosis in cancer. Nat Cancer 2020;1:755–7.
[19] Cifci D, Foersch S, Kather JN. Artificial intelligence to identify genetic alterations in conventional histopathology. J Pathol 2022;257(4):430–44. https://doi.org/10.1002/path.5898.
[20] Ghaffari Laleh N, Ligero M, Perez-Lopez R, Kather JN. Facts and hopes on the use of artificial intelligence for predictive immunotherapy biomarkers in cancer. Clin Cancer Res 2022. https://doi.org/10.1158/1078-0432.CCR-22-0390.
[21] Brockmoeller S, Echle A, Ghaffari Laleh N, Eiholm S, Malmstrøm ML, Plato Kuhlmann T, et al. Deep learning identifies inflamed fat as a risk factor for lymph node metastasis in early colorectal cancer. J Pathol 2022;256:269–81.
[22] Kiehl L, Kuntz S, Höhn J, Jutzi T, Krieghoff-Henning E, Kather JN, et al. Deep learning can predict lymph node status directly from histology in colorectal cancer. Eur J Cancer 2021;157:464–73.
[23] Meier A, Nekolla K, Hewitt LC, Earle S, Yoshikawa T, Oshima T, et al. Hypothesis-free deep survival learning applied to the tumour microenvironment in gastric cancer. Hip Int 2020;6:273–82.
[24] Saillard C, Dubois R, Tchita O, Loiseau N, Garcia T, Adriansen A, et al. Blind validation of MSIntuit, an AI-based pre-screening tool for MSI detection from histology slides of colorectal cancer. bioRxiv 2022. https://doi.org/10.1101/2022.11.17.22282460.
[25] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55–63.
[26] Metzger M-L, Behrens H-M, Böger C, Haag J, Krüger S, Röcken C. MET in gastric cancer–discarding a 10% cutoff rule. Histopathology 2016;68:241–53.
[27] Grosser B, Glückstein M-I, Dhillon C, Schiele S, Dintner S, VanSchoiack A, et al. Stroma AReactive Invasion Front Areas (SARIFA) - a new prognostic biomarker in gastric cancer related to tumor-promoting adipocytes. J Pathol 2022;256:71–82.
[28] Ghaffari Laleh N, Muti HS, Loeffler CML, Echle A, Saldanha OL, Mahmood F, et al. Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. Med Image Anal 2022;79:102474.
[29] Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, et al. A method for normalizing histology slides for quantitative analysis. 2009 IEEE Int Symp Biomed Imaging: From Nano to Macro 2009:1107–10.
[30] Wang X, Du Y, Yang S, Zhang J, Wang M, Zhang J, et al. RetCCL: clustering-guided contrastive learning for whole-slide image retrieval. Med Image Anal 2023;83:102645. https://doi.org/10.1016/j.media.2022.102645.
[31] Saldanha OL, Loeffler CML, Niehues JM, van Treeck M, Seraphin TP, Hewitt KJ, et al. Self-supervised deep learning for pan-cancer mutation prediction from histopathology. bioRxiv 2022. https://doi.org/10.1101/2022.09.15.507455.
[32] Seraphin TP, Luedde M, Roderburg C, van Treeck M, Schneider P, Buelow RD, et al. Prediction of heart transplant rejection from routine pathology slides with self-supervised Deep Learning. medRxiv 2022. https://doi.org/10.1101/2022.09.29.22279995.
[33] Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. arXiv 2018 Available at: ⟨http://arxiv.org/abs/1802.04712⟩.
[34] Niehues JM, Quirke P, West NP, Grabsch HI, van Treeck M, Schirris Y, et al. Generalizable biomarker prediction from cancer pathology slides with self-supervised deep learning: a retrospective multi-centric study. Cell Rep Med 2023;4(4):100980. https://doi.org/10.1016/j.xcrm.2023.100980.
[35] Kleppe A. Area under the curve may hide poor generalisation to external datasets. ESMO Open 2022;7:100429.
[36] Röcken C. Predictive biomarkers in gastric cancer. J Cancer Res Clin Oncol 2023;149:467–81.
[37] Wulczyn E, Steiner DF, Moran M, Plass M, Reihs R, Tan F, et al. Interpretable survival prediction for colorectal cancer using deep learning. NPJ Digit Med 2021;4:71.
[38] Martin B, Grosser B, Kempkens L, Miller S, Bauer S, Dhillon C, et al. Stroma AReactive Invasion Front Areas (SARIFA)-A new easily to determine biomarker in colon cancer-results of a retrospective study. Cancers 2021;13(19):4880. https://doi.org/10.3390/cancers13194880.
[39] Reitsam NG, Märkl B, Dintner S, Sipos E, Grochowski P, Grosser B, et al. Alterations in natural killer cells in colorectal cancer patients with Stroma AReactive Invasion Front Areas (SARIFA). Cancers 2023;15. https://doi.org/10.3390/cancers15030994.
[40] Szalai L, Jakab Á, Kocsmár I, Szirtes I, Kenessey I, Szijártó A, et al. Prognostic Ability of tumor budding outperforms poorly differentiated clusters in gastric cancer. Cancers 2022;14. https://doi.org/10.3390/cancers14194731.
[41] Ueno H, Ishiguro M, Nakatani E, Ishikawa T, Uetake H, Murotani K, et al. Prognostic value of desmoplastic reaction characterisation in stage II colon cancer: prospective validation in a Phase 3 study (SACURA Trial). Br J Cancer 2021;124:1088–97.
[42] Wang X, Chen Y, Gao Y, Zhang H, Guan Z, Dong Z, et al. Predicting gastric cancer outcome from resected lymph node histopathology images using deep learning. Nat Commun 2021;12:1637.
[43] Huang B, Tian S, Zhan N, Ma J, Huang Z, Zhang C, et al. Accurate diagnosis and prognosis prediction of gastric cancer using deep learning on digital pathological images: a retrospective multicentre study. EBioMedicine 2021;73:103631.