nature medicine

ARTICLES

https://doi.org/10.1038/s41591-022-01709-2

Deep learning-enabled assessment of cardiac allograft rejection from endomyocardial biopsies

Jana Lipkova[1,2,3,16], Tiffany Y. Chen[1,2,3,16], Ming Y. Lu[1,2,3,4], Richard J. Chen[1,2,3,5], Maha Shady[1,2,3,5], Mane Williams[1,2,3,5], Jingwen Wang[1,6], Zahra Noor[1], Richard N. Mitchell[1,7], Mehmet Turan[8], Gulfize Coskun[8], Funda Yilmaz[9], Derya Demir[9], Deniz Nart[9], Kayhan Basak[10], Nesrin Turhan[10], Selvinaz Ozkara[10], Yara Banz[11], Katja E. Odening[12,13] and Faisal Mahmood[1,2,3,14,15] ✉

Endomyocardial biopsy (EMB) screening represents the standard of care for detecting allograft rejections after heart transplant. Manual interpretation of EMBs is affected by substantial interobserver and intraobserver variability, which often leads to inappropriate treatment with immunosuppressive drugs, unnecessary follow-up biopsies and poor transplant outcomes. Here we present a deep learning-based artificial intelligence (AI) system for automated assessment of gigapixel whole-slide images obtained from EMBs, which simultaneously addresses detection, subtyping and grading of allograft rejection. To assess model performance, we curated a large dataset from the United States, as well as independent test cohorts from Turkey and Switzerland, which includes large-scale variability across populations, sample preparations and slide scanning instrumentation. The model detects allograft rejection with an area under the receiver operating characteristic curve (AUC) of 0.962; assesses the cellular and antibody-mediated rejection type with AUCs of 0.958 and 0.874, respectively; detects Quilty B lesions, benign mimics of rejection, with an AUC of 0.939; and differentiates between low-grade and high-grade rejections with an AUC of 0.833. In a human reader study, the AI system showed non-inferior performance to conventional assessment and reduced interobserver variability and assessment time. This robust evaluation of cardiac allograft rejection paves the way for clinical trials to establish the efficacy of AI-assisted EMB assessment and its potential for improving heart transplant outcomes.

Cardiac failure is a leading cause of hospitalization in the United States and the most rapidly growing cardiovascular condition globally[1,2]. For patients with end-stage heart failure, transplantation is often the only viable solution[3]. Cardiac allograft transplantation is associated with significant risk of rejection[4]. To reduce the incidence of rejection, patients receive individually tailored immunosuppressive regimens after transplantation. Despite the medications, cardiac rejection remains the most common and serious complication, as well as the main cause of mortality in post-transplantation patients[5–8].

Given that early stages of rejections may be asymptomatic[8], patients undergo surveillance EMBs that typically start days to weeks after transplantation. Although there is no standard schedule, most centers perform frequent biopsies for 1–2 years. Thereafter, screening is center specific or only performed if clinically necessary. The gold standard for EMB evaluation consists of manual histological examination of tissue slides[3]. EMB assessment includes detection and subtyping of rejection as acute cellular rejection (ACR), antibody-mediated rejection (AMR) or concurrent cellular–antibody rejections, in addition to the identification of Quilty B lesions, which are benign mimickers of rejections. The severity of the rejection is further characterized by grade. The rejection subtype and grade govern treatment regimen and patient management. Despite several revisions to the official guidelines, the interpretation of EMBs remains challenging with limited interobserver and intraobserver reproducibility[9–11]. Overestimation of rejection can lead to increased patient anxiety, overtreatment and unnecessary follow-up biopsies, whereas underestimation may lead to delays in treatment and ultimately to worse outcomes.

Deep learning-based, objective and automated assessment of EMBs can help to mitigate these challenges, potentially improving reproducibility and transplant outcomes. Multiple studies have demonstrated the potential of AI models to reach performance comparable or even superior to that of human experts in various diagnostic tasks[12–24]. Previous attempts to algorithmically assess EMBs are limited to small datasets of manually extracted regions of interest (ROIs) or handcrafted features, did not focus on all tasks involved in EMB assessment and lacked rigorous international validation across different patient populations[25–28].

In this Article, we present Cardiac Rejection Assessment Neural Estimator (CRANE), a deep learning approach for cardiac allograft rejection screening in H&E-stained whole-slide images (WSIs). CRANE addresses all major diagnostic tasks: rejection detection, subtyping, grading and detection of Quilty B lesions. CRANE is trained with thousands of gigapixel WSIs using patient-level labels, supporting seamless scalability to large datasets without the burden

[1]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [2]Broad Institute of Massachusetts Institute of Technology (MIT) and Harvard, Cambridge, MA, USA. [3]Dana-Farber Cancer Institute, Boston, MA, USA. [4]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA. [5]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [6]Department of Computer Science, University of California San Diego (UCSD), La Jolla, CA, USA. [7]Harvard-MIT Health Sciences and Technology (HST), Cambridge, MA, USA. [8]Institute of Biomedical Engineering, Bogazici University, Istanbul, Turkey. [9]Faculty of Medicine, Department of Pathology, Ege University, Izmir, Turkey. [10]Department of Pathology, University of Health Sciences, Ankara, Turkey. [11]Institute of Pathology, University of Bern, Bern, Switzerland. [12]Department of Cardiology, Inselspital, Bern University Hospital, Bern, Switzerland. [13]Institute of Physiology, University of Bern, Bern, Switzerland. [14]Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [15]Harvard Data Science Initiative, Harvard University, Cambridge, MA, USA. [16]These authors contributed equally: Jana Lipkova, Tiffany Y. Chen. ✉e-mail: faisalmahmood@bwh.harvard.edu

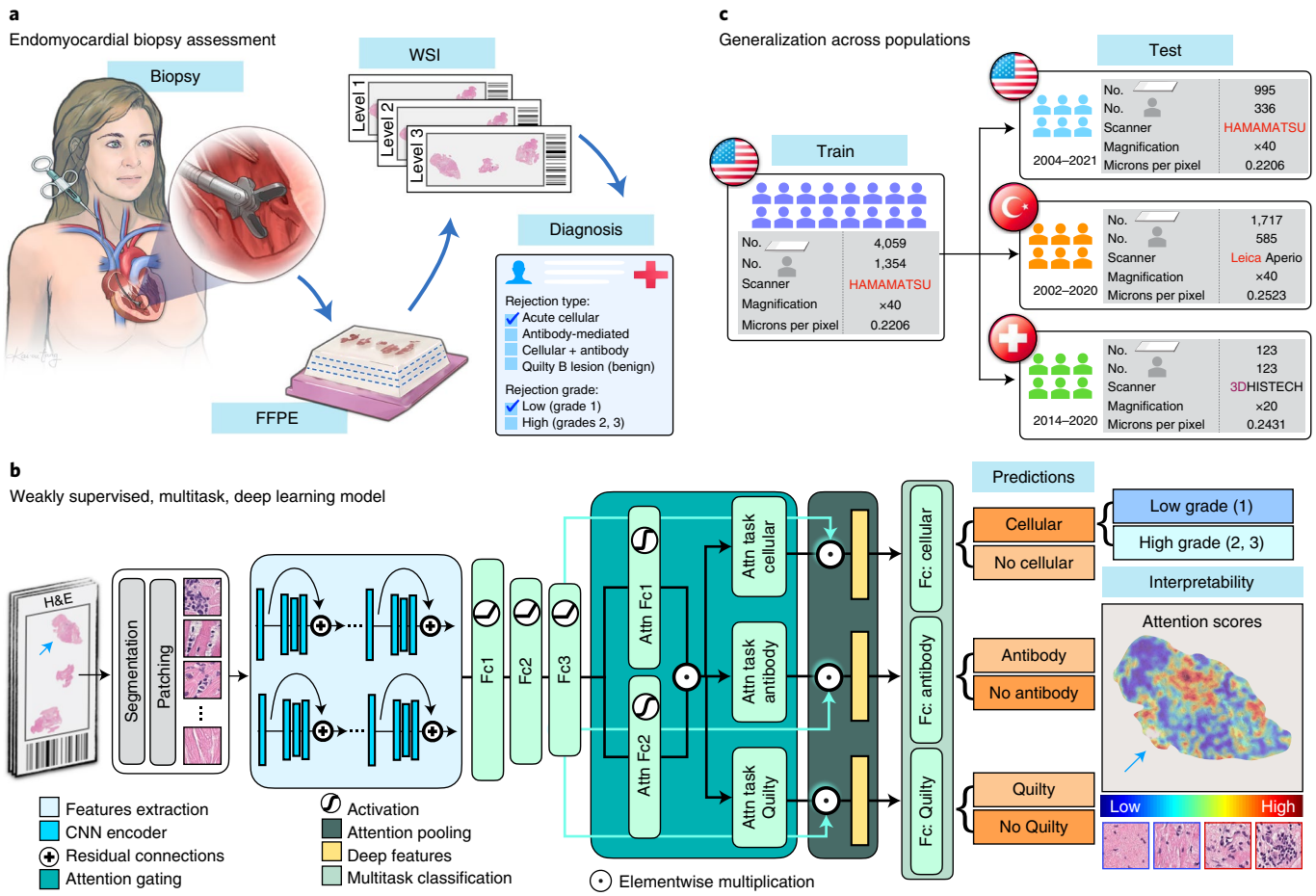


Fig. 1 | CRANE workflow. a, Fragments of endomyocardial tissue are formalin fixed and paraffin embedded (FFPE). Each paraffin block is cut into slides with three consecutive levels and stained with H&E. Each slide is digitized and serves as an input for the model. **b**, CRANE first segments tissue regions in the WSI and patches them into smaller sections. A pretrained CNN encoder is used to extract features from the image patches, which are further fine-tuned through a set of three fully connected layers, marked as Fc1, Fc2 and Fc3. A weakly supervised, multitask, multilabel network was constructed to simultaneously identify normal tissue and different rejection conditions (cellular, antibody and/or Quilty B lesions). A separate classifier was trained to estimate the rejection grade. The attention (Attn) scores, reflecting the relevance of each image region toward the model prediction, can be visualized in the form of whole-slide attention heatmaps. **c**, The model was trained on the US cohort, using 70% of cases for training and 10% for validation and model selection. Evaluation of the model was conducted on the internal US dataset using the remaining 20% cases as hold-out set and two external cohorts from Turkey and Switzerland. These datasets are analyzed in detail in Supplementary Table 1.

of manual annotations. The model performance is evaluated on three test cohorts from the United States, Turkey and Switzerland, using different biopsy protocols and scanner instrumentation. For model interpretability and introspection, visual representation of the model predictions is obtained with high-resolution heatmaps, reflecting the diagnostic relevance of morphological regions within the biopsy. An independent reader study is performed to assess the model's consensus with manual expert assessment and to demonstrate the potential of AI assistance in reducing interobserver variability and assessment time. An overview of the study is shown in Fig. 1.

Results

Endomyocardial biopsy assessment via deep learning. CRANE was developed on an internal dataset consisting of 5,054 gigapixel WSIs from 1,690 patient biopsies collected from Brigham and Women's Hospital. Each biopsy had three WSIs representing three levels of the tissue block (Fig. 1a). Each biopsy had associated labels for the presence of rejection, characterization of the rejection type as cellular and/or antibody-mediated, rejection grade and presence of Quilty B lesion. All cases diagnosed as AMRs were confirmed with C4d immunohistochemical (IHC) staining. The internal dataset was used for model training, validation and hold-out testing. To rigorously assess model adaptability, CRANE was tested on two additional independent international cohorts sourced from Turkey (1,717 WSIs from 585 patients) and Switzerland (123 WSIs from 123 patients). These independent international test sets were deliberately constructed to reflect the high data variability present across populations and medical centers, as they used different biopsy protocols, slide preparation and staining mechanisms and scanner vendors, which are all known contributors to image variability[17,18,29]. The variation in the color distribution across cohorts is illustrated in Extended Data Fig. 1. A specification of the data collection protocols and the patient cohorts are provided in Fig. 1, Supplementary Table 1 and the Dataset description section of the Methods.

CRANE is a high-throughput, multitask framework that simultaneously detects ACR, AMR and Quilty B lesions, including their concurrent appearances using H&E-stained WSIs as the input (Fig. 1b). Owing to the large size of gigapixel histology images, it is computationally inefficient and usually not possible to apply deep learning models directly to the WSI. To circumvent this, we first automatically segment tissue regions and patch them into smaller sections.

Using transfer learning, a deep residual convolutional neural network (CNN) is deployed as an encoder to extract low-dimensional features from raw image patches. The extracted features are further tuned to histology-specific representations through a fully connected neural network. This allows for a high-dimensional gigapixel WSI to be embedded into a set of compact low-dimensional feature vectors for efficient training and inference. The feature vectors serve as input for the attention-based multiple instance learning (MIL) module[17,30]. Using the patient diagnoses for the supervision, the attention module learns to rank the relative importance of each biopsy region in the determination of each classification task. The parameters in the first attention layers are shared among all three tasks to enable the identification of atypical myocardial tissue. Subsequently, a separate branch is used for each task, allowing the model to identify morphology specific to each diagnosis. The feature representations from all tissue patches, weighted by their respective predicted attention score for each task, are aggregated in attention pooling[30]. The resulting slide-level features are then evaluated by the corresponding task-specific classifier, which independently determines the presence or absence of ACR, AMR and Quilty B lesions. The presence of overall rejection is obtained by combining the ACR and AMR predictions for each biopsy. A separate, single-task MIL classifier is trained to determine the grade for each detected rejection, discriminating between low-grade (that is, grade 1R for ACRs and grades pAMR-1i and pAMR-1h for AMRs) and high-grade (that is, grade 2R and 3R for ACRs and grades pAMR2 and pAMR3 for AMRs) cases. The high-grade rejections are then further differentiated into the specific subgrades using a separate network. Owing to the rare occurrence of grade 3 rejections (9 cases in the US cohort, 13 in the Turkish cohort and none in the Swiss cohort), this problem cannot be addressed at the whole-slide level (that is, using the weakly supervised approach). Instead, we design a supervised model that uses pixel-level annotations. This model, summarized in Extended Data Fig. 2, is trained on patches extracted from the biopsy regions corresponding to the specific rejection grades as annotated by experts. Additional details on the model architecture and hyperparameters are given in the Methods.

Evaluation of model performance. The internal US dataset is partitioned into groups for training (70%), validation (10%) and hold-out testing (20%). The partition is constructed in a way to ensure a balanced proportion of each diagnosis across all groups. Multiple slides from the same biopsy are always presented in the same group. The model training is performed at the slide level, where each slide is considered an independent data point with patients' diagnoses as labels.

We examine model performance across different magnifications, including ×40, ×20 and ×10. These experiments show that the rejections and Quilty B lesions are best estimated by fusion of model predictions from multiple magnifications (using averaging of the prediction scores), whereas the best performance for grade prediction was achieved at ×10 magnification (Extended Data Fig. 3a,b). This observation is consistent with manual assessment, wherein signs of rejections usually need to be examined at different magnifications, whereas lower magnification is more informative for grade determination because it provides more context on the extent of myocardial injury. Learning curves for all tasks are shown in Extended Data Fig. 3c.

For the hold-out US test set, which was not used during training, the model accuracy (ACC) and AUC at the patient level are as follows: cellular rejections, AUC = 0.958 (95% confidence interval (CI) 0.940–0.977) and ACC = 0.893; antibody-mediated rejection, AUC = 0.874 (95% CI 0.801–0.946) and ACC = 0.899; Quilty B lesions, AUC = 0.939 (95% CI 0.910–0.969) and ACC = 0.920; and grade, AUC = 0.833 (95% CI 0.764–0.901) and ACC = 0.818. Performance for overall rejection detection reached an AUC of 0.962 (95% CI 0.943–0.980) and an ACC of 0.899. The patient-level performance is reported in Fig. 2 and Supplementary Table 2, and slide-level scores are given in Extended Data Fig. 4 and Supplementary Tables 3 and 4. The high performance in detecting overall rejections implies the model's potential for screening negative cases. CRANE is also able to differentiate between ACRs and their benign mimickers, Quilty B lesions. Other benign injuries such as old biopsy sites, focal healing injuries and tissue scars can often imitate the appearance of ACRs, particularly in the early stages after transplantation[31]. Although the model was not explicitly trained to account for such injuries, it learned to differentiate between ACRs and injuries; with an AUC of 0.983 (95% CI 0.950–1.00) for early injuries (that is, 6 months after transplant), and AUC of 0.977 (95% CI 0.963–0.991) for all (early and old) injuries (Supplementary Table 5). The model identifies AMRs using only conventional H&E-stained slides, including pAMR-1i rejections, which are clinically diagnosed by IHC status. However, owing to rare appearance of pAMR-1i rejections (with only six cases in all cohorts together), it is not possible to conclude if the model is able to identify the IHC-specific predictive features from the H&E-slides[32,33]. Compared with other tasks, model performance is lower for rejection grading, similar to the trends observed in manual EMB assessment[9–11]. This can be attributed to the higher complexity of the grading task over rejection detection, as grade is characterized by the type and extent of the tissue injury. The high-grade ACRs were further differentiated into the respective grades (that is, grade 2 versus 3) through the supervised grading model (see the Methods), which achieved an AUC of 0.929 (95% CI 0.861–0.997) and an ACC of 0.885 at the patch level (Extended Data Fig. 2). For slide-level predictions, obtained by the fusion of the patch-level estimates through a majority voting, AUC was 0.960 and ACC was 0.800 (Supplementary Table 6).

The performance of the model is comparable for slide-level and patient-level predictions. These results further imply that the patient-level labels, which are readily available in clinical records, are sufficient for model training without the burden of assigning diagnoses for each slide separately. Although there may be label noise in the training data due to observer variability, deep learning models are known to be robust to a substantial amount of label noise[34]. The model performance across different patient groups (Supplementary Tables 7–9) further illustrates the model robustness to variations in patient age, gender and time since transplant. The model robustness can be also assessed through the confidence of the predictions, as shown in Extended Data Fig. 5.

Generalization to independent international test cohorts. The AI model is applied, without any form of domain adaptation, stain normalization or model tuning, to the two independent international external cohorts (from Turkey and Switzerland). To stress test the trained model, we ensured that these cohorts reflect the large-scale variability in the training cohort, including different slide preparation protocols and slide scanning vendors. Consistent with our assessment on the US hold-out dataset, the multitask model is applied to all available magnification downsamples in each cohort, and the grade is assessed from the ×10 magnification. The patient-level scores are reported in Fig. 2, Extended Data Fig. 6 and Supplementary Table 2, and the slide-level performance is reported in Extended Data Fig. 4 and Supplementary Tables 3 and 4. The scores for the refinement of the high-grade cellular rejections are reported in Extended Data Fig. 2 and Supplementary Table 6. Adapting the model from internal to external cohorts led to a drop in performance that varied between 0.02 and 0.13 for AUC and 0.02 and 0.14 for ACC scores, similar to other deep learning models when applied to external independent datasets[16,18,35]. The decrease in performance can be attributed to the large data variations among the cohorts. As demonstrated by previous studies, small changes such as rescanning the same slides with a different scanner and variation in the slide preparation can also reduce performance,

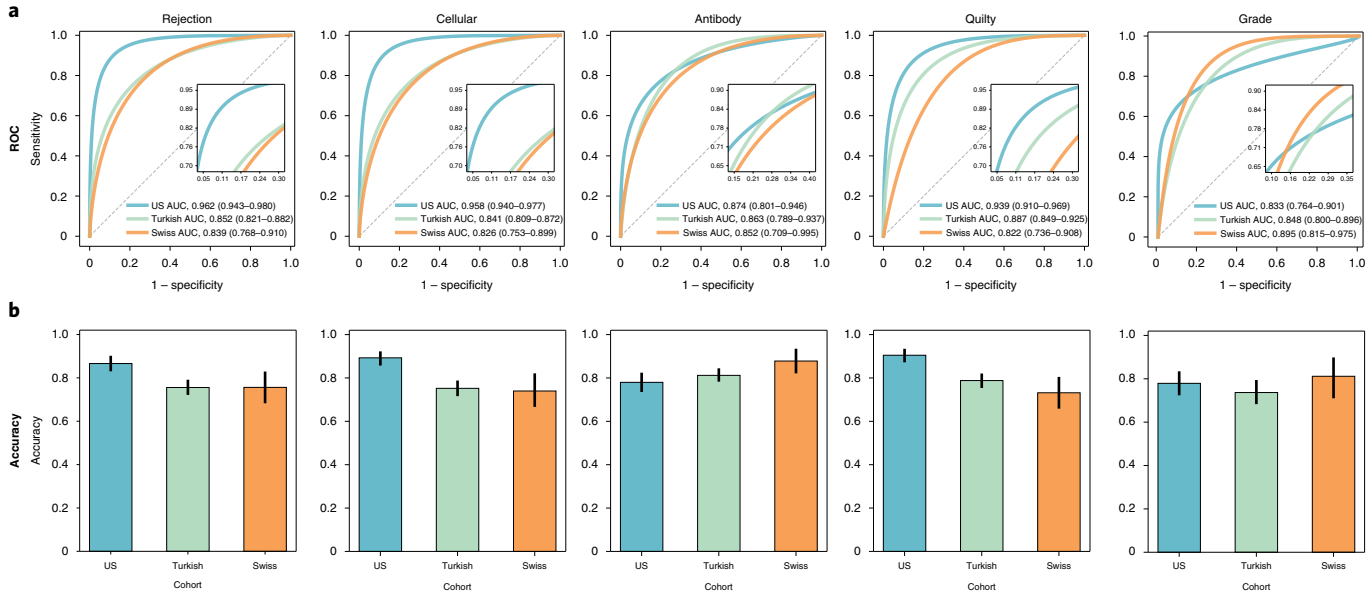


Fig. 2 | Performance of the CRANE model at patient level. The CRANE model was evaluated on the test set from the United States (995 WSIs, 336 patients) and two independent external cohorts from Turkey (1,717 WSIs, 585 patients) and Switzerland (123 WSIs, 123 patients). **a**, ROC curves for the multitask classification of EMB evaluation and grading at the patient level. The AUC is reported together with the 95% CIs for each cohort. **b**, The bar plots reflect model accuracy for each task. Error bars (black lines) indicate 95% CIs, and the center is always the computed value for each cohort.

even on simple binary classification tasks[18]. Furthermore, different staining techniques, slide-thickness, micron-per-pixel variations across scanners (even for the same resolution[17,29]) and signal-to-noise ratio are factors that are known to affect model performance and were all present in our cohorts. Despite the decrease in performance on the external dataset, the fact that the model reached comparable scores on two highly distinct international cohorts without any form of domain adaptation implies that the model has the potential to generalize across large variations presented in histopathology data. Assessment of model performance on patient subgroups (Supplementary Tables 7–9) further shows the model robustness to variations in patient age, gender and time since transplant, which is especially interesting as the Turkish dataset includes also cases of pediatric patients.

Model interpretability. To visualize and interpret the model predictions, we generated attention heatmaps for each task by mapping the normalized attention scores to their corresponding spatial location in the WSI, as shown in Fig. 3 and in our public interactive demonstration (http://crane.mahmoodlab.org). Technical details on heatmap generation are provided in the Methods. The attention scores represent the model's interpretation of the diagnostic relevance of each biopsy region. Although no manual annotations were used for training, the model has learned to discriminate between atypical and normal myocardial tissue. In all tasks, the regions with high attention scores typically correspond to diagnostically relevant morphology, whereas the low attention scores indicate benign tissue. However, the attention maps should not be overinterpreted as segmentation of rejection regions. They merely reflect the relative importance of each biopsy region (relative to others) for the model predictions. As such, we observed that the model pays higher attention to the regions with atypical tissue, which might also include tissue without sign of rejection, such as benign injuries or artifacts. However, the model predictions are derived from the regions with the highest attention scores for each task. This is further illustrated in Extended Data Fig. 7, which shows the attention heatmaps for a case with concurrent ACR, AMR and Quilty B lesion. Although at the slide-level view the attention maps appear relatively similar for all three tasks, the regions with the highest attention scores correspond to tissue with the task-specific morphology. This example also provides a visual representation of the model's ability to differentiate between ACR and similarly appearing Quilty B lesions. There is a visible difference in the heatmap appearance between the rejection and grading tasks; the rejection task heatmap is more refined and detailed, and the grading task heatmap has coarser morphology. This observation is consistent with manual EMB examination, where fine morphological details are used to determine the rejection type, whereas grade evaluation requires more spatial context. The regions with high attention scores provide useful insights into the morphologies driving the model predictions. Further quantitative analysis shows a close agreement between the model and expert assessment of diagnostically relevant regions (Extended Data Fig. 8).

Attention heatmaps for the external datasets are presented in Fig. 4. In all cohorts, the model strongly attended to regions with rejection morphology, whereas the low attention scores corresponded to regions with normal or benign myocardial tissue. This further demonstrates the model's ability to generalize across diverse populations and different scanners. To further investigate the model predictions and limitations, we performed a detailed analysis of failure cases. The top misclassified cases from each cohort, specifically the cases in which the model made incorrect predictions with the highest confidence, are detailed in Fig. 5. We observed that for all failure cases, the model has correctly assigned high attention scores to the regions with rejection morphology. However, the detection of the rejection regions does not necessarily warrant correct predictions. It simply means that the model has considered these regions to be the most relevant diagnostically; the corresponding features might not be discriminative enough for the model to draw a correct prediction. Despite the incorrect predictions, the identified regions with high attention scores can guide pathologists toward the relevant biopsy regions and in turn reduce interobserver and intraobserver variability.

Comparison with human readers. To compare the performance of the AI model with that of expert pathologists, we recruited five certified pathologists (mean experience, 10.5 years). Previous studies

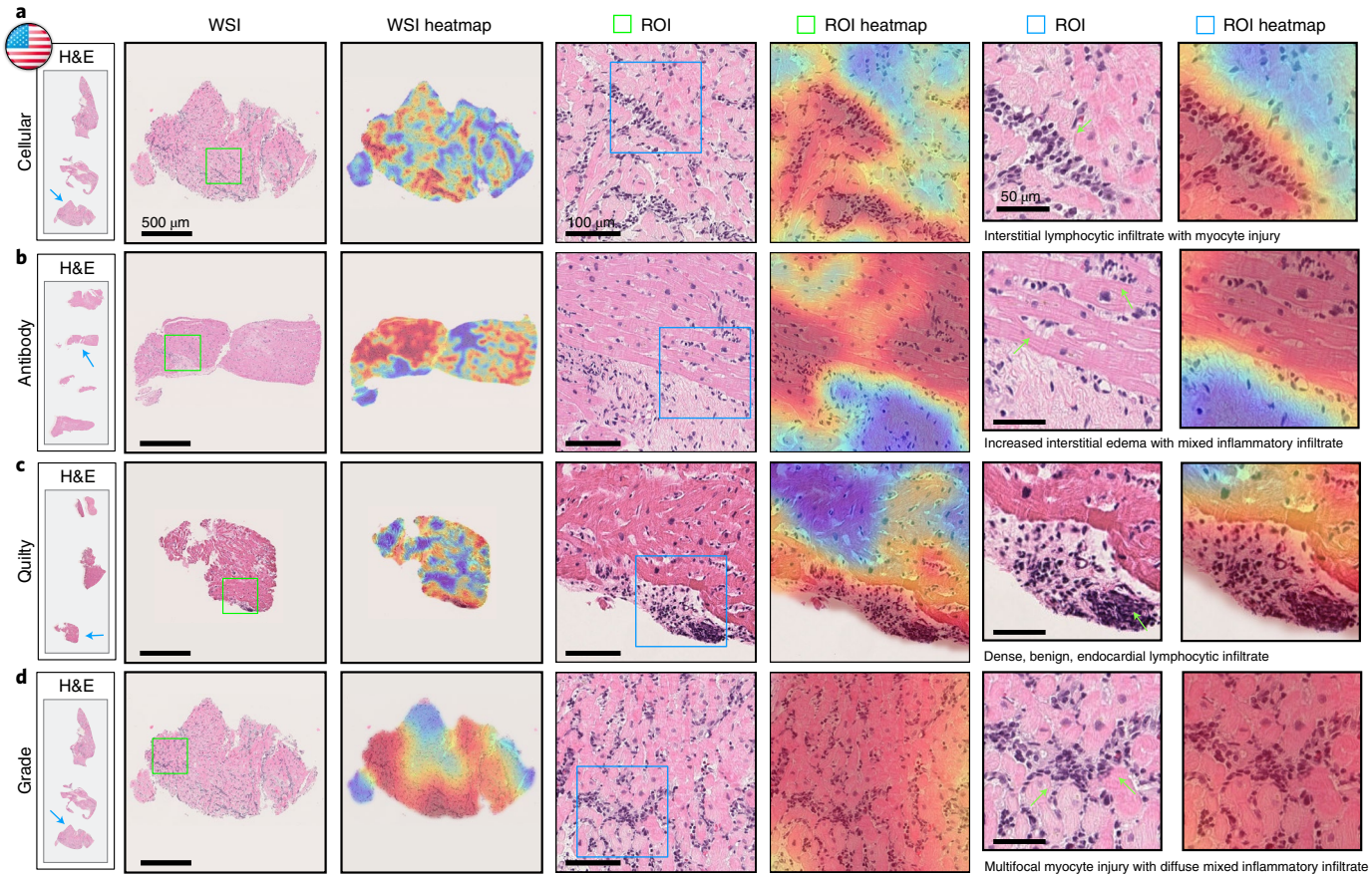


Fig. 3 | Visualization of the attention heatmaps for the US cohort. A sample of WSIs of EMBs with different diagnoses are shown in the first column. The second column displays a magnified view of the cardiac tissue samples indicated by the blue arrows in the first column. An attention heatmap corresponding to each slide was generated by computing the attention scores for each predicted diagnosis (third column). The fourth column shows a magnified view of the ROIs indicated by the green squares in the second column, and the corresponding attention heatmaps are displayed in the fifth column. The last two columns show a magnified view of the ROIs indicated by the blue squares together with the corresponding attention heatmap. Green arrows indicate specific morphology corresponding to the textual description. The colormap of the attention heatmaps ranges from red (high attention) to blue (low attention). **a**, Cellular rejection. The highest attention scores correspond to regions with increased interstitial lymphocytic infiltrates and associated myocyte injury, whereas the adjacent, lower attention scores indicate healthier myocytes without injury. **b**, Antibody-mediated rejection. The highest attention scores correspond to regions with edema within the interstitial spaces in addition to increased mixed inflammatory infiltrate, comprised of eosinophils, neutrophils and lymphocytes. The adjacent lower attention scores indicate background fibrosis, stroma and healthier myocytes. **c**, Quilty B lesion. The highest attention scores identify a single, benign lymphocytic focus within the endocardium, without injury or damage to the myocardium. The lower attention scores indicate background and healthy myocytes. **d**, Cellular grade. The highest attention scores indicate diffuse, prominent interstitial lymphocytic infiltrates with associated myocyte injury, representing severe rejection. The lower attention regions indicate background fibrosis and unaffected, healthier myocytes.

assessing the performance of AI models against that of pathologists for cancer diagnosis have relied on building a reference standard based on the consensus of human experts[15,16,19]. However, as the variability in EMB assessment is substantially larger than that in cancer diagnosis[11], the AI model easily identified cases in which a clear consensus could be reached. In light of this, we investigated the interobserver variability among experts and the variability between the AI system and individual experts. We randomly selected a set of 150 EMBs from the Turkish cohort, including 91 ACRs, 23 AMRs (including 14 concurrent ACR and AMR cases) and 50 normal biopsies. All experts were blinded to pathology reports and previous assessments made on these biopsies. The cases were presented to the pathologists in random order through a digital web-based platform, and no timing constraint was imposed on the readers. For evaluation purposes, the pathologists were asked to derive the diagnosis from H&E slides only, mimicking CRANE's setup without the use of IHC for AMR detection. We first calculated the interobserver agreement between each pair of pathologists using Cohen's kappa (κ), which accounts for agreement by chance, and then calculated the agreement between individual expert pathologists and the CRANE model trained on US cases. The agreement between the five experts was comparable to the findings of previous studies that assessed interobserver agreement in EMBs[10]. For all tasks, we observed that the AI predictions are not inferior to the human reads (Extended Data Fig. 9). For example, the average agreement for rejection detection between individual pathologists was $\kappa = 0.537$ (moderate agreement), whereas the average agreement between individual pathologists and CRANE was $\kappa = 0.639$ (substantial agreement).

To further evaluate the potential of CRANE to serve as an assistive diagnostic tool, we conducted an additional reader study to assess the benefit of showing model-generated heatmaps during case assessment. The pathologists were asked to assess 150 EMBs from the Turkish cohort. The readers were randomly assigned to one of two groups: one to make assessments on WSIs only and the second to make assessments on WSIs with AI assistance in the form of attention heatmaps shown as a semitransparent layer at the top of

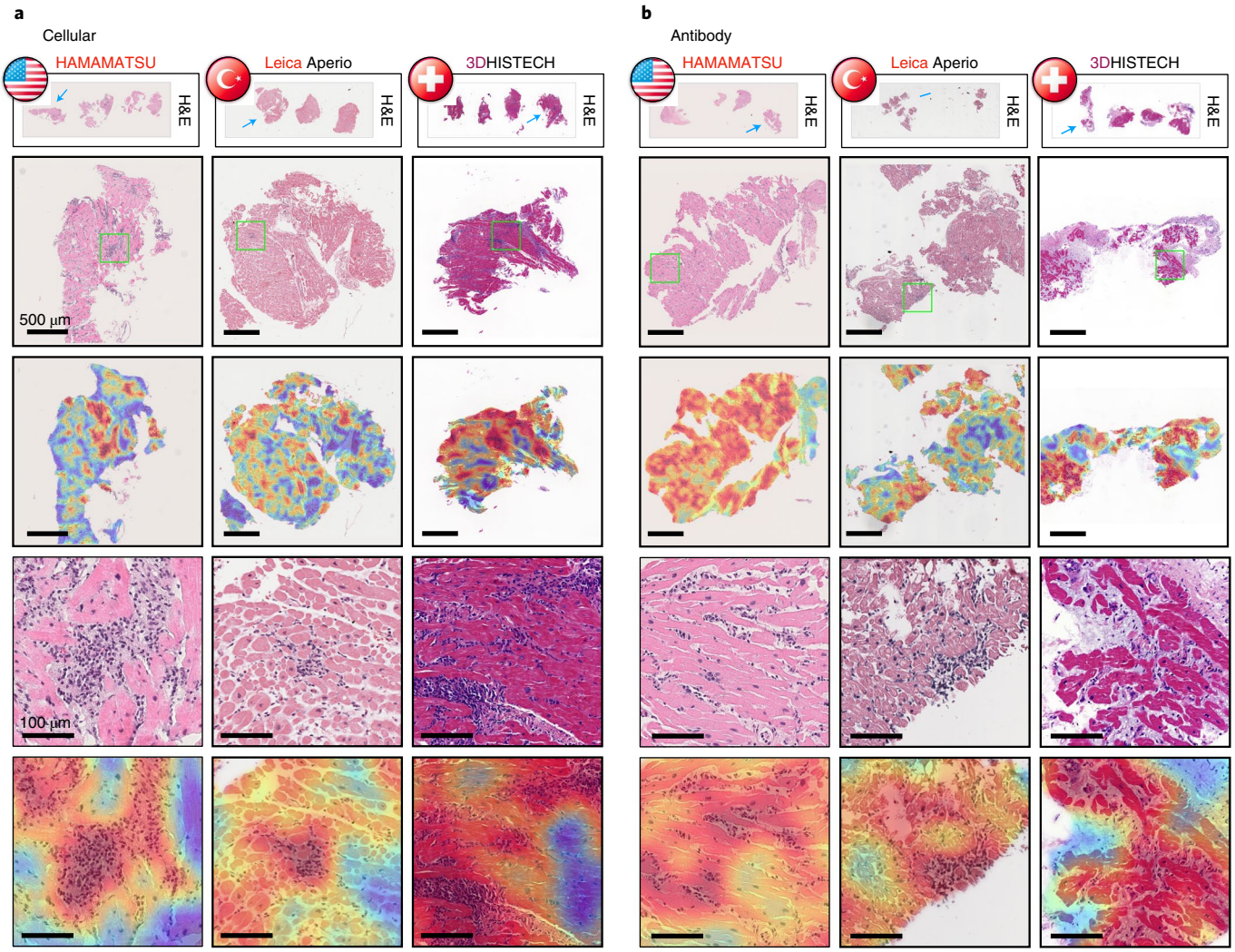


Fig. 4 | Analysis of attention heatmaps in the three independent test cohorts. a,**b**, Cases with cellular (**a**) and antibody-mediated (**b**) rejections sourced from the US (left columns), Turkish (middle columns) and Swiss (right columns) cohorts. The types of scanner used for each cohort are indicated. The first row shows WSIs from each corresponding center, and the second row shows magnified views of cardiac tissue samples indicated by the blue arrows in the first row. The corresponding attention heatmaps are shown in the third row. The colormap of the attention heatmaps ranges from red (high attention) to blue (low attention). The last two rows show a magnified view of ROIs marked by the green squares in the second row along with the corresponding attention heatmaps.

the H&E slide. All features available in our public interactive demonstration were available to the readers, and no timing constraint was imposed on the readers. Following a 4-week washout period, the pathologists repeated the task, but the readers from the first group used WSIs and AI assistance, whereas the second group used only WSIs. For the purpose of this evaluation, we created ground truth labels based on the consensus of the readers from the interobserver variability experiment presented in Extended Data Fig. 9. Although this reader study is small and limited in scope, it suggests that using CRANE as an assisting tool increased the accuracy for all tasks and reduced the assessment time for all readers (Extended Data Fig. 10). These results indicate the potential of AI assistance in reducing interobserver variability and increasing the efficiency of manual biopsy assessment, and set the stage for large-scale studies and clinical trials to assess the benefit of AI-assisted EMB assessment.

Discussion

Here we present CRANE, a weakly supervised deep learning model for automated screening of EMBs in H&E-stained WSIs. Using multitask learning, the model can simultaneously identify ACRs, AMRs, Quilty B lesions and their concurrent appearance, while an additional network is used to determine the rejection grade. The model has been trained using only patient diagnoses readily available in the clinical records, allowing seamless model deployment for large datasets from multiple centers. The usability of CRANE has been demonstrated on two independent international test cohorts, which reflect diverse geographic populations, scanners and biopsy protocols.

Although the datasets used in all three cohorts reflected the screening population of each participating institution, the proportion of AMR cases was considerably lower than that of ACR cases in all cohorts, which is one limitation of the study. This can be attributed to the rarer occurrence of AMRs and to the later recognition of this rejection type by the medical community. Although CRANE's performance for the rejection grading is comparable to that of human experts, this task remains challenging and will benefit from future model improvements.

Our model demonstrates the promise of integrating AI into the diagnostic workflow. However, optimal use of weakly supervised

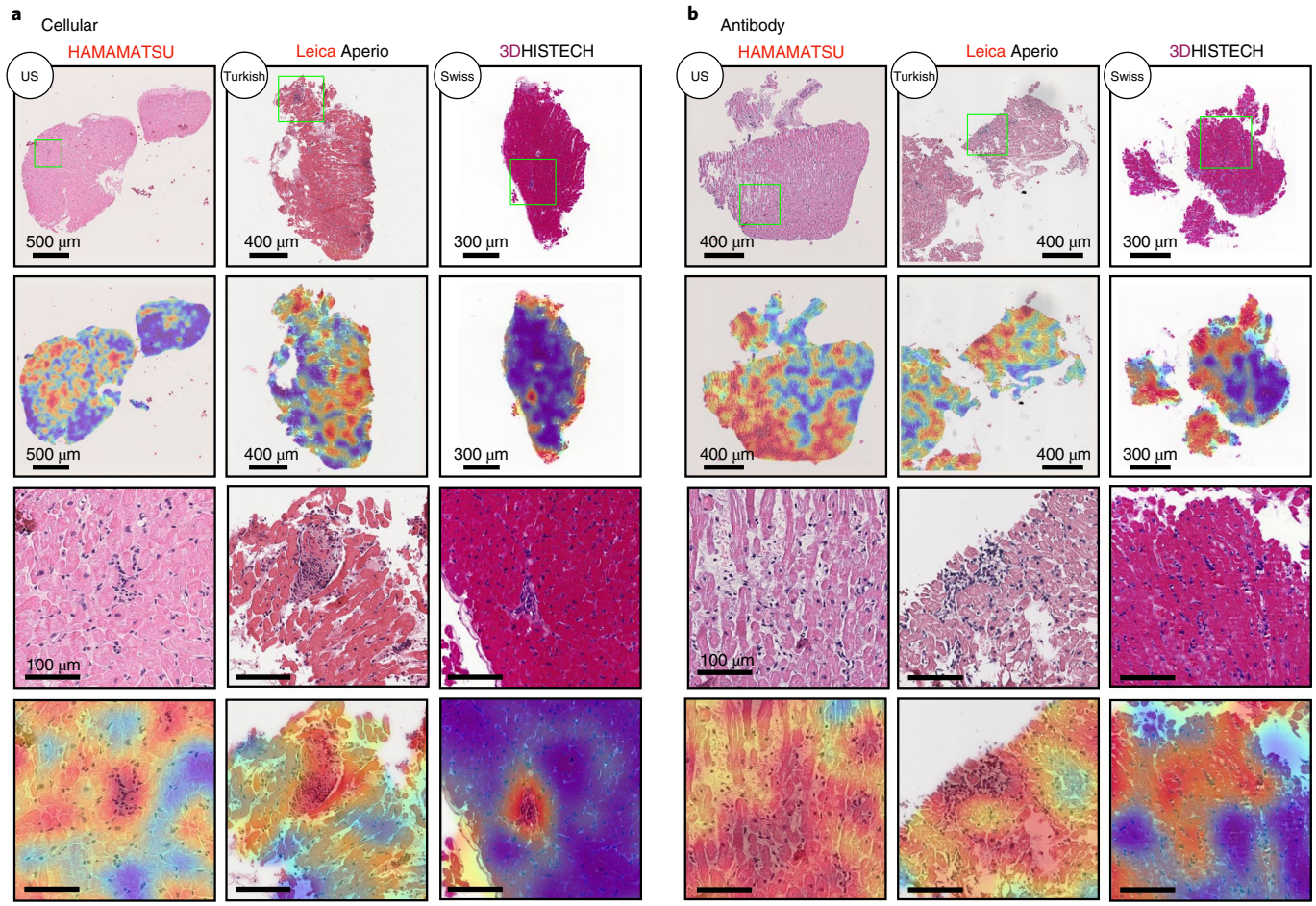


Fig. 5 | Analysis of failure cases using attention heatmaps in the three independent test cohorts. a,**b**, Cases with cellular (**a**) and antibody-mediated (**b**) rejections that were incorrectly identified as normal by the model. Results are shown for the US (left columns), Turkish (middle columns) and Swiss (right columns) cohorts. The types of scanner used for each cohort are indicated. The WSIs from each center and the corresponding attention heatmaps are shown in the first and second rows, respectively. The third row shows magnified views in the ROIs marked by the green squares, and the accompanying attention heatmaps are shown in the last row.

models in clinical practice remains to be determined. The specific advantages of CRANE suggest its potential to act as an assistive diagnostic tool with aims to increase the efficiency of EMB assessment and decrease interobserver variability by highlighting predictive regions; such assistive tools that highlight areas of interest for human analysis are currently in use for cytology[36]. Improved robustness and accuracy of rejection assessment could reduce the number of unnecessary follow-up biopsies, a highly valuable outcome given the cost and risks associated with EMBs. CRANE can be further deployed to automatically screen for critical and time-sensitive cases that may benefit from priority inspections.

Although our study focuses on morphology-based biopsy assessment based on current standards of the International Society for Heart and Lung Transplantation (ISHLT), future works could benefit from the integration of clinical end points such as echocardiography or cardiac hemodynamic measurements to improve patient stratification. Additional incorporation of emerging molecular biomarkers such as donor-specific antibodies, intragraft mRNA transcripts[31,37–39], cell-free DNA[40], exosomes[41] and gene expression profiling[42] could further enhance our understanding of the pathophysiology of cardiac rejection and involved immune interactions. This study lays the foundation and prompts the need for future prospective trials to fully evaluate the extent to which AI-based assessments can improve heart transplant outcomes.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-022-01709-2.

Received: 15 January 2021; Accepted: 19 January 2022;
Published online: 21 March 2022

References

1. Ziaeian, B. & Fonarow, G. C. Epidemiology and aetiology of heart failure. *Nat. Rev. Cardiol.* **13**, 368–378 (2016).
2. Benjamin, E. J. et al. Heart disease and stroke statistics–2018 update: a report from the American Heart Association. *Circulation* **137**, e67–e492 (2018).
3. Badoe, N. & Shah, P. in *Contemporary Heart Transplantation* (eds Bogar, L. & Stempien-Otero, A.) 3–12 (Springer, 2020).
4. Orrego, C. M., Cordero-Reyes, A. M., Estep, J. D., Loebe, M. & Torre-Amione, G. Usefulness of routine surveillance endomyocardial biopsy 6 months after heart transplantation. *J. Heart Lung Transplant.* **31**, 845–849 (2012).
5. Lund, L. H. et al. The Registry of the International Society for Heart and Lung Transplantation: thirty-fourth adult heart transplantation report—2017; focus theme: allograft ischemic time. *J. Heart Lung Transplant.* **36**, 1037–1046 (2017).
6. Colvin-Adams, M. & Agnihotri, A. Cardiac allograft vasculopathy: current knowledge and future direction. *Clin. Transplant.* **25**, 175–184 (2011).

7. Kfoury, A. G. et al. Cardiovascular mortality among heart transplant recipients with asymptomatic antibody-mediated or stable mixed cellular and antibody-mediated rejection. *J .Heart Lung Transplant.* **28**, 781–784 (2009).
8. Costanzo, M. R. et al. The International Society of Heart and Lung Transplantation Guidelines for the care of heart transplant recipients. *J. Heart Lung Transplant.* **29**, 914–956 (2010).
9. Kobashigawa, J. A. The search for a gold standard to detect rejection in heart transplant patients: are we there yet? *Circulation* **135**, 936–938 (2017).
10. Angelini, A. et al. A web-based pilot study of inter-pathologist reproducibility using the ISHLT 2004 working formulation for biopsy diagnosis of cardiac allograft rejection: the European experience. *J. Heart Lung Transplant.* **30**, 1214–1220 (2011).
11. Crespo-Leiro, M. G. et al. Concordance among pathologists in the second Cardiac Allograft Rejection Gene Expression Observational Study (CARGO II). *Transplantation* **94**, 1172–1177 (2012).
12. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
13. Bejnordi, B. E. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).
14. Ouyang, D. et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **580**, 252–256 (2020).
15. Chen, P.-H. C. et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nat. Med.* **25**, 1453–1457 (2019).
16. McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
17. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
18. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
19. Bulten, W. et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol.* **21**, 233–241 (2020).
20. Chen, R. J. et al. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans. Med. Imaging* https://doi.org/10.1109/TMI.2020.3021387 (2020).
21. Mahmood, F. et al. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE Trans. Med. Imaging* **39**, 3257–3267 (2020).
22. Fu, Y. et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat. Cancer* **1**, 800–810 (2020).
23. Kather, J. N. et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat. Cancer* **1**, 789–799 (2020).
24. Lu, M. Y. et al. AI-based pathology predicts origins for cancers of unknown primary. *Nature* **594**, 106–110 (2021).
25. Peyster, E. G. et al. An automated computational image analysis pipeline for histological grading of cardiac allograft rejection. *Eur. Heart J.* **42**, 2356–2369 (2021).
26. Tong, L., Hoffman, R., Deshpande, S. R. & Wang, M. D. Predicting heart rejection using histopathological whole-slide imaging and deep neural network with dropout. In *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)* 1–4 (IEEE, 2017).
27. Nirschl, J. J. et al. A deep-learning classifier identifies patients with clinical heart failure using whole-slide images of H&E tissue. *PLoS ONE* **13**, e0192726 (2018).
28. Peyster, E. G., Madabhushi, A. & Margulies, K. B. Advanced morphologic analysis for diagnosing allograft rejection: the case of cardiac transplant rejection. *Transplantation* **102**, 1230–1239 (2018).
29. Sellaro, T. L. et al. Relationship between magnification and resolution in digital pathology systems. *J. Pathol. Inform.* **4**, 21 (2013).
30. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. In *Proc. 35th International Conference on Machine Learning* (eds Dy, J. & Krause, A.) 2132–2141 (PMLR, 2018).
31. Halloran, P. F. et al. Exploring the cardiac response to injury in heart transplant biopsies. *JCI Insight* **3**, e123674 (2018).
32. Schmauch, B. et al. A deep learning model to predict RNA-seq expression of tumours from whole slide images. *Nat. Commun.* **11**, 3877 (2020).
33. Coudray, N. et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
34. Karimi, D., Dou, H., Warfield, S. K. & Gholipour, A. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. *Med. Image Anal.* **65**, 101759 (2020).
35. Mitani, A., Hammel, N. & Liu, Y. Retinal detection of kidney disease and diabetes. *Nat. Biomed. Eng.* **5**, 487–489 (2021).
36. Biscotti, C. V. et al. Assisted primary screening using the automated ThinPrep Imaging System. *Am. J. Clin. Pathol.* **123**, 281–287 (2005).
37. Halloran, P. F. et al. Building a tissue-based molecular diagnostic system in heart transplant rejection: the heart Molecular Microscope Diagnostic (MMDx) System. *J. Heart Lung Transplant.* **36**, 1192–1200 (2017).
38. Duong Van Huyen, J.-P. et al. MicroRNAs as non-invasive biomarkers of heart transplant rejection. *Eur. Heart J.* **35**, 3194–3202 (2014).
39. Giarraputo, A. et al. A changing paradigm in heart transplantation: an integrative approach for invasive and non-invasive allograft rejection monitoring. *Biomolecules* **11**, 201 (2021).
40. De Vlaminck, I. et al. Circulating cell-free DNA enables noninvasive diagnosis of heart transplant rejection. *Sci. Transl. Med.* **6**, 241ra77 (2014).
41. Kennel, P. J. et al. Serum exosomal protein profiling for the non-invasive detection of cardiac allograft rejection. *J. Heart Lung Transplant.* **37**, 409–417 (2018).
42. Anglicheau, D. & Suthanthiran, M. Noninvasive prediction of organ graft rejection and outcome using gene expression patterns. *Transplantation* **86**, 192–199 (2008).

Methods

Dataset description. *US cohort.* The model was developed using data from the US cohort collected at Brigham and Women's Hospital, which comprised 5,054 WSIs from 1,690 internal EMBs (2004–2021). No statistical methods were used to determine sample size. For rare subtypes, grades and concurrent appearances the sample size was limited by the number of cases available in-house or off-site pathology archives. For classes with abundant slides, deep learning models were trained while increasing the number of input slides until asymptotic improvement of the model performance was achieved to suggest an appropriate sample size was obtained. The study involved retrospective analysis of pathology slides and patients were not directly involved or recruited for the study. The Mass General Brigham institutional review board approved the retrospective analysis of pathology slides and corresponding pathology reports (protocol 2020P000234). Informed consent was waived for analyzing archival pathology slides retrospectively. All pathology slides were deidentified before scanning and digitization. All digital data, including WSIs, pathology reports and EMRs were deidentified before computational analysis and model development. We collected all available ACR cases between 2007 and 2021, and AMR cases between 2011 and 2021. As the number of high-grade rejections has decreased over the past decade, we included all grade 2 and 3 cellular rejections from 2004 onwards. As a majority of EMBs are diagnosed as normal, these cases were collected from 2017 to 2020, wherein the number of normal cases were selected to be approximately equivalent to the number of rejection cases. Having a higher amount of normal cases might not be beneficial, as machine-learning methods tend to develop a bias toward the majority class present in the training data[43]. The diagnoses for all cases were distilled from deidentified pathology reports, which have been determined by expert cardiac pathologists based on the contemporary ISHLT criteria (that is, the ISHLT-2004 guideline for ACRs and the ISHLT-2013 guideline for AMRs). Due to the changes in the ISHLT guidelines over the collecting period, all AMR cases collected before 2013 and all ACR cases collected before the 2004 revision have been re-evaluated to ensure compatibility with the contemporary ISHLT guidelines. The IHC staining was performed for all of the cases included in the study as part of the clinical biopsy assessment. The only exception is a set of three high-grade cellular rejections from 2004, for which the C4d IHC status is not reported. Given that we cannot guarantee the absence of pAMR-1i (only immunopathologic findings present) rejection in these three cases, they were not considered when evaluating the model performance for AMRs. The slides were scanned using a HAMAMATSU S210 scanner at ×40 (and included ×20 and ×10 pyramid downsamples) . The dataset is randomly partitioned and stratified by diagnosis into a training set (70% of cases), a validation set (10% of cases) and a hold-out test set (20% of cases). The partitioning was performed at the patient level to ensure that all of the slides from the same biopsy were always placed into the same set. Patient demographics (age, gender and time after transplant) and their distribution are reported in Supplementary Table 1.

The model was additionally evaluated on two international independent test sets submitted from medical centers in Turkey and Switzerland. The institutional review board at Ege University, Turkey, approved the retrospective analysis of pathology slides (protocol 20-11 T/61). The study was considered exempt from review board approval in Switzerland according to the Swiss Human Research Act (HFG 810.30, 30 September 2011). Informed consent was also waived for the study locations in Turkey and Switzerland for analyzing archival pathology slides retrospectively. All data from Turkey and Switzerland were deidentified at their corresponding institutions and received at Brigham and Women's Hospital for model testing. The diagnoses for these cohorts, determined by expert cardiac pathologists, were extracted from the deidentified pathology reports at each institution. C4d IHC staining was used to confirm all AMRs at both centers. The slides from different cohorts were prepared at their respective institutions using diverse biopsy protocols, staining approaches and scanner vendors. Although all three centers use similar biopsy protocols, their slide preparation and scanning processes differ. For example, in the US and Turkish cohorts, each biopsy level is stored on a separate slide, whereas in the Swiss cohort all three biopsy levels are placed in a single slide. As such, there is no difference in the patient-level and slide-level predictions for the Swiss cohort.

Turkish cohort. The Turkish cohort comprised 1,717 WSIs from 585 patients, collected between 2002 and 2020, at Ege University Hospital, using the following data-selection protocol. We attempted to collect data for all possible classes and combinations. Sample size for rare subtypes, grades and concurrant occurrences was limited by the number of cases available at the Ege University pathology archives. The cellular rejections were collected from 2006 to 2020, and the antibody-mediated rejections were collected from 2011 to 2020, including all available cases. Additionally, all available grade 2R and 3R rejections from 2002 to 2005 were also included in the test set to increase the number of high-grade rejections in the test set. Due to the concurrent appearance of cellular and antibody-mediated rejections, the dataset contains an additional four AMRs cases from 2007. The non-rejection cases were collected from 2014 to 2020; the amount of non-rejection cases was selected to be proportional to the overall number of collected rejection cases. To maintain compatibility with contemporary ISHLT-grading guidelines, all cases from the early years have been re-evaluated by pathologists at Ege Universtiy Hospital. IHC staining was performed on all cases as part of the clinical assessment. The only exception is a set of 26 patients for which the IHC status is not reported. Similar to the US cohort, these cases were not considered when evaluating model performance for antibody-mediated rejections. The slides were scanned using a Leica Aperio CS2 scanner at ×40 and included a ×10 pyramid downsample. Patient demographics (age, gender and time after transplant) and their distribution are reported in Supplementary Table 1.

Swiss cohort. The Swiss cohort comprised 123 WSIs from 123 patients, collected between 2014 and 2020 at Bern University Hospital (following the ISHLT-2004 guideline for ACRs and the ISHLT-2013 guideline for AMRs). The patient selection aimed to reflect the variation in the population. The slides were scanned at ×20 with a ×10 pyramid downsample using a 3DHISTECH 250 Flash scanner. Despite the lower magnification used for the Swiss cohort, the number of microns per pixel of the WSIs is comparable to that of images from the other two cohorts. Due to the small number of cardiac allograft transplants performed in Switzerland (on average less than 40 cases per year, based on data from the Global Observatory on Donation and Transplantation[44]), information relating to patient demographics (age, gender and time after transplant) was not provided to preserve patients' privacy, additional data distribution details are provided in Supplementary Table 1.

Weakly supervised multitask model architecture. To assess the state of EMB, it is necessary to determine whether the biopsy tissue is normal or demonstrates signs of rejection. In the latter case, it is necessary to determine whether the rejection is acute cellular, antibody-mediated or a benign Quilty B lesion mimicking cellular rejection. The different rejection states are not mutually exclusive but can appear concurrently. The extent of the rejection is further quantified by the rejection grade. CRANE uses two weakly supervised networks to address all of these tasks. The first network assesses the state of the EMB, and the second network determines the grade of the detected rejections, discriminating between low-grade (that is, grade 1R for ACRs and grades pAMR-1i and pAMR-1h for AMRsgrade 1) and high-grade (that is, grade 2R and 3R for ACRs and grades pAMR2 and pAMR3 for AMRs grades 2 and 3) rejections. An auxiliary model is designed to refine the inferred high-grade rejections into grades 2 and 3.

CRANE uses attention mechanism, which as previously shown[17,18] is particularly suitable for dealing with WSIs. Although the architecture of CRANE is similar to that of previous approaches dealing with cancer detection[17,24], cardiac rejections can often exhibit substantially large variability in disease presentation. Specific rejection types can be represented by several different morphologies (for example, diffusive versus focal inflammatory infiltrates). At the same time, similarly appearing patterns can encode various rejection types or their mimickers, and multiple rejection types can coexist within a single slide. Predictions made by CRANE can be obtained on both the patient level and slide level. For simplicity, we describe the model's architecture using the slide-level formalism, and the adaptation to the patient-level predictions is provided afterward.

To minimize the number of models required for the EMB evaluation, we opted for a multitask architecture that allows simultaneous prediction of the presence or absence of acute cellular rejection (task 1), antibody-mediated rejection (task 2) and Quilty B lesions (task 3). In this formulation, all acute cellular rejections of non-zero grades (that is, grades 1R, 2R and 3R) are considered as a single class. The same applies for all the non-zero grades of antibody-mediated rejections (that is, grade pAMR-1i, pAMR-1h, pAMR2, pAMR2 and pAMR3). All rejection grades were combined into a single class because different rejection types are characterized by diverse morphology, and thus the role of the multitask model is to learn to distinguish among the rejection types, their benign mimickers and healthy tissue. Posing the problem as three multitask binary classifiers allows the model to make all three predictions simultaneously while also enabling analysis of the biopsy regions responsible for predictions of each task via attention heatmaps.

Given that histology images are typically composed of several gigapixels, it is computationally inefficient—and often infeasible—to apply CNN models directly on the entire WSI. Standard deep learning models generally use smaller ROIs for model training. However, this approach requires manual annotation of representative image regions in each WSI for each considered task. The high cost and human bias of manual annotations often limits the deployment of such deep learning models on a large scale. To overcome these limitations, we use the MIL approach. In the MIL formalism, a WSI is considered as a collection (referred to as 'bag') of smaller image regions (referred to as 'instances'). Using the weak supervision in the form of patient diagnosis, the model learns to identify which image regions are representative for the given diagnosis.

Preprocessing. To preprocess the WSI data, we used the publicly available CLAM WSI-analysis toolbox[17]. First, the tissue regions in each biopsy slide were segmented and stored as object contours. To train models at different magnifications, we extracted non-overlapping 256 × 256-sized image patches from each magnification using the segmentation contours. For computational efficiency, each raw image patch is embeded into low-dimensional feature representation vector using a ResNet50[45] convolutional neural network pretrained on ImageNet[46]. Specifically, using spatial average pooling after the third residual block leads to a single 1,024-dimensional feature vector representation for each image patch. For

a WSI represented as a bag of K instances (patches), we denote the patch-level embedding corresponding to patch k as $\mathbf{z}_k \in \mathbb{R}^{1,024}$

Histology-specific features. To enable the model to learn histology-specific feature representations, the deep features extracted by the ResNet50 encoder are further tuned through three stacked fully connected layers, Fc_1, Fc_2 and Fc_3, parameterized by $W_1 \in \mathbb{R}^{768\times1,024}$, $\mathbf{b}_1 \in \mathbb{R}^{768}$; $W_2 \in \mathbb{R}^{512\times768}$, ; $\mathbf{b}_2 \in \mathbb{R}^{512}$;and $W_3 \in \mathbb{R}^{512\times512}$, $\mathbf{b}_3 \in \mathbb{R}^{512}$ respectively, each followed by rectified linear unit (ReLU) activation. In this way, each patch feature embedding $\mathbf{z}_k \in \mathbb{R}^{1,024}$ is mapped to a 512-dimensional vector $\mathbf{h}_k$ defined as:

$$\mathbf{h}_k = \mathrm{ReLU}\left[W_3\left(\mathrm{ReLU}\left\{W_2\left[\mathrm{ReLU}\left(W_1\mathbf{z}_k + \mathbf{b}_1\right)\right] + \mathbf{b}_2\right\} + \mathbf{b}_3\right)\right]. \quad (1)$$

The low-dimensional feature embeddings $\{\mathbf{h}_k\}_{k=1}^{K}$ serve as input for the multitask attention pooling module.

Multitask attention pooling. The attention module aggregates information from all tissue regions and learns to rank the relative importance of each region toward the determination of each classification task. All three tasks are learned jointly, sharing the model parameters in the first layer, while a separate branch is used for each learning objective in the second layer. In this architecture, the first attention layer consists of two parallel attention networks, Attn-Fc_1 and Attn-Fc_2, with weight parameters $V_a \in \mathbb{R}^{384\times512}$ and $U_a \in \mathbb{R}^{384\times512}$ (shared across all tasks), and one independent layer $W_{a,t} \in \mathbb{R}^{1\times384}$ for each task t. The attention module is trained to assign an attention score $a_{k,t}$ for each patch k and task t, given as:

$$a_{k,t} = \frac{\exp\left\{W_{a,t}\left[\tanh\left(V_a\mathbf{h}_k\right) \odot \mathrm{sigma}\left(U_a\mathbf{h}_k\right)\right]\right\}}{\sum_{j=1}^{K}\exp\left\{W_{a,t}\left(\tanh\left(V_a\mathbf{h}_j\right) \odot \mathrm{sigma}\left(U_a\mathbf{h}_j\right)\right)\right\}}. \quad (2)$$

For simplicity, the bias parameters are omitted in equation (2). After applying the softmax activation, the attention scores reflect the relevance of each image region toward determining the given diagnosis, where the highly relevant regions have scores close to 1 and the diagnostically non-specific regions have scores close to 0. The representation of the entire slide for the given task t, $\mathbf{h}_{\mathrm{slide},t}$, is then computed by averaging feature representations of all patches in the given slide, weighted by their respective attention scores $a_{k,t}$ as follows:

$$\mathbf{h}_{\mathrm{slide},t} = \sum_{k=1}^{K} a_{k,t}\,\mathbf{h}_k. \quad (3)$$

Multitask classifier. The deep features $\mathbf{h}_{\mathrm{slide},t}$ of each task are fed into the final classification layer. This layer consists of three binary classifiers, one per task. The slide-level probability prediction score for each task t is computed as:

$$p_t = \mathrm{softmax}\left(W_{\mathrm{cls},t}\,\mathbf{h}_{\mathrm{slide},t} + \mathbf{b}_{\mathrm{cls},t}\right), \quad (4)$$

where each task-specific classification layer (cls) is parametrized by $W_{\mathrm{cls},t} \in \mathbb{R}^{2\times512}$, $\mathbf{b}_{\mathrm{cls},t} \in \mathrm{R}^2$.

Rejection grade. The rejection grade is estimated by the second weakly supervised MIL network. The network takes WSI as input, which was classified as rejection by the multitask model, and performs binary classification that discriminates between low-grade (grade 1) and high-grade (grades 2 and 3) cases. The network has the same architecture as the multitask model previously described, with the only difference being that the model for grading is posed as a single-task binary classifier.

Hyperparameters and training details. We randomly sampled slides using a minibatch size of one WSI and used multitask learning to supervise the neural network during training. For each slide, the total loss is a weighted sum of loss functions from all three tasks:

$$\mathcal{L}_{\mathrm{total}} = c_1\mathcal{L}_{\mathrm{cls},1} + c_2\mathcal{L}_{\mathrm{cls},2} + c_3\mathcal{L}_{\mathrm{cls},3}. \quad (5)$$

The standard cross-entropy was used for all tasks and $c_1 = c_2 = c_3 = 1/3$. In the single-task model for the grade prediction, standard cross-entropy is used as well, but without scaling parameter c. The model parameters are updated through the Adam optimizer with a learning rate of 2×10^{-4} and ℓ_2-weight decay of 1×10^{-5}. The running averages of the first and second moment of the gradient were computed with the default coefficient values (that is, $\beta_1 = 0.9$, $\beta_2 = 0.999$). The epsilon term for the numerical stability is also used with the default value of 1×10^{-8}. To protect the model from potential overfitting, dropout layers with $P = 0.25$ are used after every hidden layer.

Model selection. During training, the model's performance is monitored for each epoch using the validation set. The model is trained for a minimum of 50 epochs and a maximum of 200 epochs. After the initial 50 epochs, if the validation loss (that is, sum of all tasks) has not decreased for 20 consecutive epochs, early stopping is triggered and the best model with the lowest validation loss is used for reporting the performance on the hold-out test sets. Fivefold cross-validation is used to further assess the robustness of the model's training, where the best model is selected based on the performance on the validation set.

Patient-level predictions. The model predictions can be obtained at both the slide level and patient level. At the slide level, each WSI is treated as an independent data point, where patient diagnosis is used as a label for each slide in the biopsy. In the patient-level approach, all slides and their corresponding patches from a biopsy are treated as a single input (unified bag) for the model. The model then aggregates information from all slides to perform the patient-level predictions.

Attention heatmaps. For each task in the multitask prediction problem, the attention scores predicted by the model can be used to visualize the relative importance assigned to each region in the WSI. Similar to how WSIs are processed for training and inference, we first tile each WSI into 256×256 patches without overlap, perform feature extraction and compute the attention score for each patch for each prediction task. For increased visual smoothness, we subsequently increase the tiling overlap up to 90% of the chosen patch size and convert the computed attention scores to percentile scores between 0.0 (low attention) and 1.0 (high attention) using the initial sets of attention scores that were computed without overlapping as references. Finally, normalized scores are mapped to their corresponding spatial location in the WSI, and scores within overlapped regions are reduced by summing and averaging. Lastly, the attention heatmap is registered with the original H&E image and displayed as a semitransparent overlay. Examples of attention maps are provided in Figs. 3 and 4 and Extended Data Fig. 7, and can also be visualized in high resolution through our interactive demonstration (http://crane.mahmoodlab.org).

Refinement of high-grade rejections. To demonstrate the feasibility of using deep learning to distinguish between grade 2 and 3 cellular rejections, we trained a separate supervised deep network. This task could not be accomplished in a weakly supervised manner because of the limited cases available for grade 3 rejections. We obtained rough, pixel-level expert annotations of representative regions for each rejection type in WSIs from 33 EMBs in the US cohort (14 cases with grade 2 rejections and 9 cases with grade 3 rejections) and 26 EMBs in the Turkish cohort (13 cases with grade 2 rejections and 13 cases with grade 3 rejections). Image patches of 512×512 (without overlap) were then extracted from the annotated regions. The US dataset was randomly partitioned into groups for training (60%), validation (20%) and hold-out testing (20%) while ensuring patches from the same patient were always drawn into the same subset. We used a pretrained CNN based on the EfficientNet-B3 architecture[47], initialized with weights pretrained on ImageNet. The model was trained on two Nvidia 2080 Ti graphics processing units (GPUs) using distributed data parallel for up to 50 epochs using a batch size of eight patches per GPU, a learning rate of 2×10^{-4} and ℓ_2-weight decay of 1×10^{-5} with the Adam optimizer with default optimizer hyperparameters (see Hyperparameters and training details). The validation loss was monitored for each epoch, and early stopping was performed when it did not improve for 10 consecutive epochs. The model checkpoint with the lowest recorded validation loss was then evaluated on the hold-out test set. The experimental setup was repeated five times, with different random partitions. The patch-level performance of the classifier is shown in Extended Data Fig. 2, and the slide-level performance is reported in Supplementary Table 6. CRANE does not address refinement of extremely rare high-grade AMRs, as there was only one case with pAMR3 grade among all three cohorts. The Swiss cohort was excluded from the analysis due to the absence of grade 3 rejection cases.

Computational hardware and software. WSIs were processed on Intel Xeon multicore central processing units (CPUs) and two consumer-grade NVIDIA 2080 Ti GPUs using the publicly available CLAM[17] whole-slide processing pipeline implemented in Python (version 3.7.5). Weakly supervised deep learning models were trained on GPUs using Pytorch (version 1.7.1). The supervised classifier for refining high-grade rejection was trained using the PytorchLightning (version 1.3.3) and timm (version 0.4.9) Python libraries. Plots were generated in Python using matplotlib (version 3.1.1) unless otherwise specified. Additionally, the following Python libraries were used for analysis and data handling: h5py (2.10.0), numpy (1.18.1), opencv-python (4.1.1), openslide-python (1.1.1), pandas (1.0.3), pillow (6.2.1), scipy (1.3.1), tensorflow (1.14.0), tesnorboardx (1.9) and torchvision (0.6). The AUC was estimated using the scikit-learn scientific computing library (version 0.22.1), based on the Mann–Whitney U statistic. Additionally, the pROC library (version 1.16.2) in R (version 3.6.1) was used for computing the 95% CIs of the true AUC using DeLong's method and binormal receiver operating characteristic (ROC) curve smoothing. The 95% CIs for accuracy scores were computed using non-parametric bootstrapping from 1,000 bootstrap samples. The interactive demonstration website was developed using OpenSeadragon (version 2.4.2) and jQuery (version 3.6.0).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Please email all requests for academic use of raw and processed data to the corresponding author. Restrictions apply to the availability of the in-house and external data, which were used with institutional permission for the current study, and are thus not publicly available. All requests will be promptly evaluated based on institutional and departmental policies to determine whether the data requested are subject to intellectual property or patient privacy obligations. Data can only be shared for non-commercial academic purposes and will require a data user agreement. A subset of whole-slide images used in the study can be accessed through the interactive demonstration available at http://crane.mahmoodlab.org. ImageNet data are available at https://image-net.org/. Source data are provided with this paper.

Code availability

All code was implemented in Python using PyTorch as the primary deep learning package. All code and scripts to reproduce the experiments of this paper are available at https://github.com/mahmoodlab/CRANE.

References

43. Dong, Q., Gong, S. & Zhu, X. Imbalanced deep learning by minority class incremental rectification. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 1367–1381 (2019).
44. Matesanz, R., Mahillo, B., Alvarez, M. & Carmona, M. Global observatory and database on donation and transplantation: world overview on transplantation activities. *Transplant. Proc.* **41**, 2297–2301 (2009).
45. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on ComputerVision and Pattern Recognition (CVPR)* 770-778 (2016).
46. Russakovsky, O. et al. ImageNet large scale visual recognition challenge.*Int. J. Comput. Vis.* **115**, 211–252 (2015).
47. Tan, M. & Le, Q. V. EfficientNet: rethinking model scaling for convolutional neural networks. In *Proc. 36th International Conference on Machine Learning* (eds Chaudhuri, K. & Salakhutdinov, R.) 6105–6114 (PMLR, 2019).

Acknowledgements

We thank A. Bruce for scanning internal cohorts of patient histology slides at Brigham and Women's Hospital (BWH); K. Bronstein, L. Cirelli and E. Askeland for querying the BWH slide database and retrieving archival slides; C. Li for assistance with EMRs and the Research Patient Data Registry (RPDR); M. Bragg, T. Mellen, S. Zimmet and T. A. Mages for logistical support; and K. Tung for anatomical illustrations. Y.B. and K.E.O. thank the Translational Research Unit team at the Institute of Pathology of the University of Bern for technical assistance and IT assistance; in particular, M. Skowronska, L. Daminescu and S. Reinhard. This work was supported in part by the BWH President's Fund, National Institute of General Medical Sciences (NIGMS) R35GM138216 (to F.M.), Google Cloud Research Grant, Nvidia GPU Grant Program and internal funds from BWH and Massachusetts General Hospital (MGH) Pathology. M.S. was supported by the National Institutes of Health (NIH) National Library of Medicine (NLM) Biomedical Informatics and Data Science Research Training Program, T15LM007092. M.W. was funded by the NIH National Human Genome Research Institute (NHGRI) Ruth L. Kirschstein National Research Service Award Bioinformatics Training Grant, T32HG002295. T.Y.C. was funded by the NIH National Cancer Institute (NCI) Ruth L. Kirschstein National Service Award, T32CA251062. R.J.C. was funded by the National Science Foundation (NSF) Graduate Fellowship. The content is solely the responsibility of the authors and does not reflect the official views of the NIGMS, NIH, NLM, NHGRI, NCI or NSF.

Author contributions

F.M., J.L. and T.Y.C. conceived the study and designed the experiments. J.L. performed the experimental analysis. T.Y.C., J.L., R.N.M., F.M. and J.W. curated training and test datasets. M.T., G.C., D.D., D.N., F.Y., K.B., N.T. and S.O. curated the Turkish independent test cohort. Y.B. and K.E.O. curated the Swiss independent test cohort. J.L., T.Y.C., M.Y.L., M.S., M.W., R.N.M., R.J.C. and F.M. analyzed the results. Z.N. created the interactive demonstration. J.L., T.Y.C. and F.M. prepared the manuscript with input and feedback from all authors. F.M. supervised the research.

Competing interests

The authors declare no competing interests.

Additional information

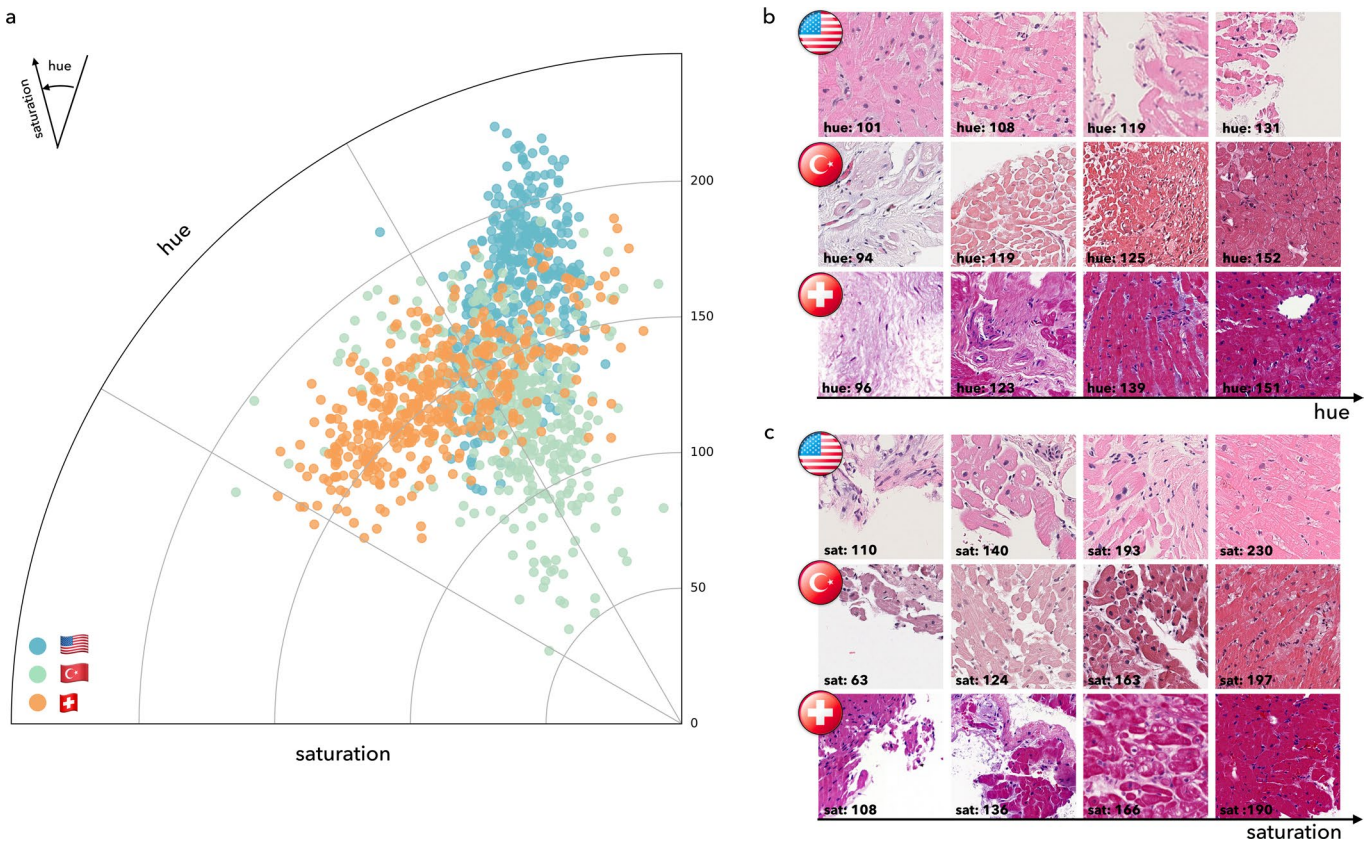
Extended data is available for this paper at https://doi.org/10.1038/s41591-022-01709-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-022-01709-2.

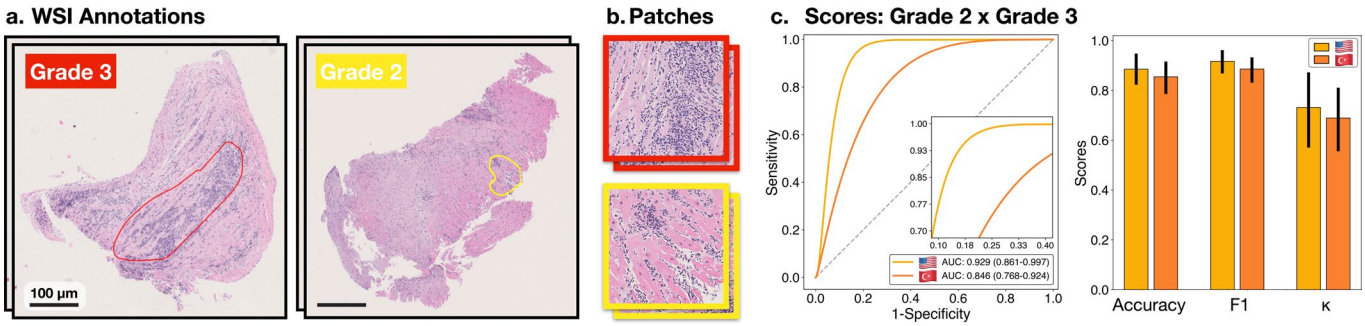
Correspondence and requests for materials should be addressed to Faisal Mahmood.

Peer review information *Nature Medicine* thanks Geert Litjens and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Michael Basson, in collaboration with the *Nature Medicine* team.

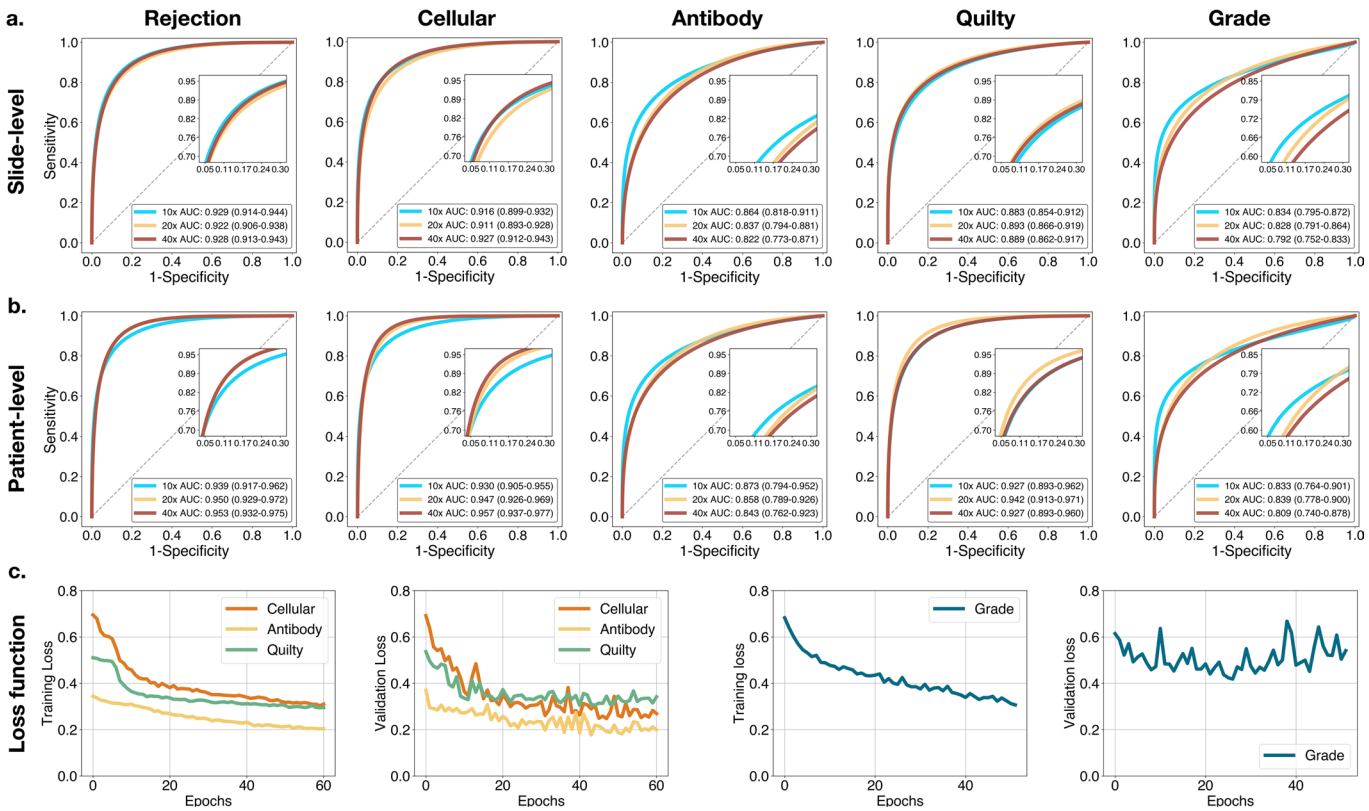
Reprints and permissions information is available at www.nature.com/reprints.



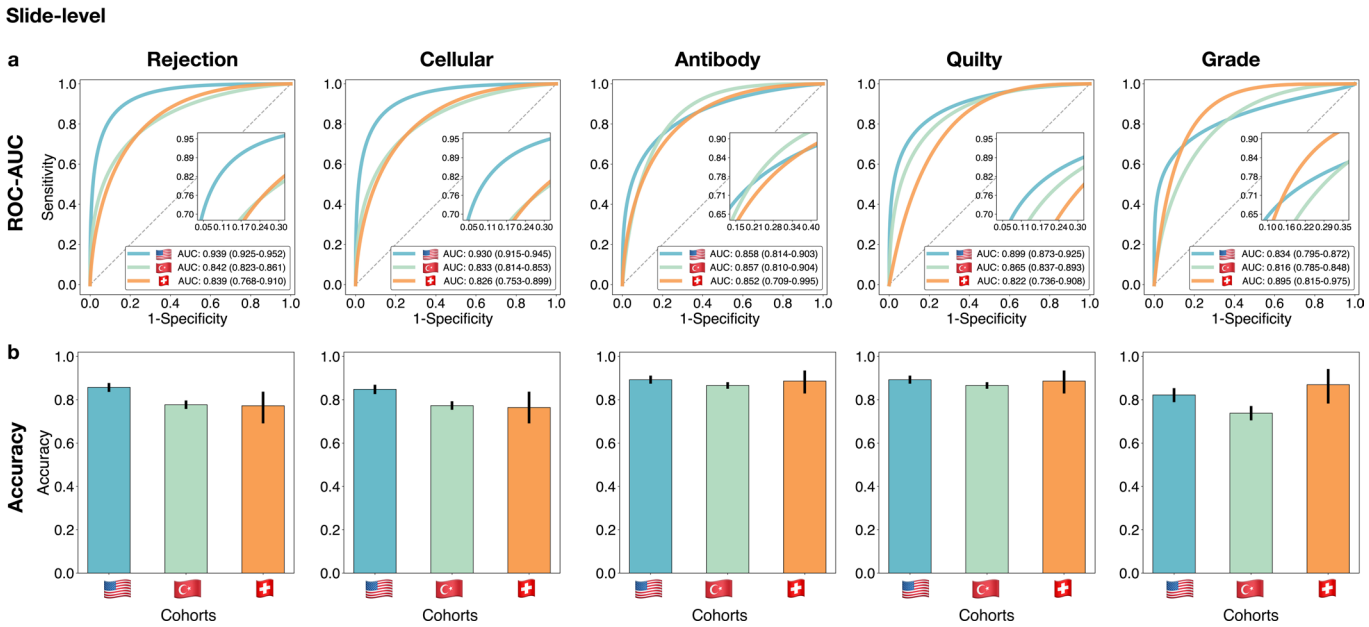
Extended Data Fig. 1 | Visualization of color distribution among the three cohorts. a. Polar scatter plot depicts the differences in the train (US) and test (US, Turkish, Swiss) cohorts, each acquired with different scanners and staining protocols. The angle represents the color (*i.e.* hue) and the polar axis corresponds to the saturation. Each point represents average hue and saturation of an image patch selected from each cohort. To construct the figure, 100 WSIs were randomly selected from each cohort. For each selected slide, 4 patches of size 1024×1024 at ×10 magnification were randomly selected from the segmented tissue regions. A hue–saturation–density color transform is taken to correct for the logarithmic relationship between light intensity and stain amount. The Swiss cohort demonstrates a large variation in both hue and saturation whereas the US and Turkish cohorts have a relatively uniform saturation but variable hue. Examples of patches with diverse hue and saturation from each cohort are shown in subplots **b**. and **c**.



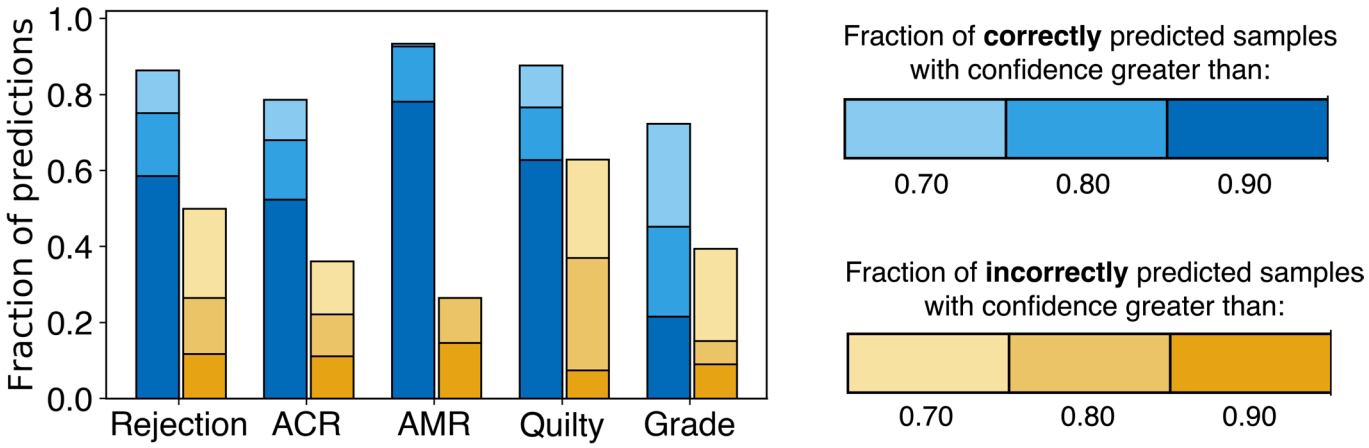
Extended Data Fig. 2 | Classification of high-grade cellular rejections. A supervised, patch-level classifier is trained to refine the detected high-grade (2 R + 3 R) cellular rejections into grades 2 and 3. A subplot **a**. shows manual annotations of the predictive region for each grade as outlined by pathologist. **b**. Patches extracted from the respective annotation regions serve as input for the binary classifier. Subplot **c**. shows the model performance at patches extracted from the US (m = 290 patches) and Turkish (m = 131 patches) cohort. Reported are ROC curves with 95% confidence intervals (CIs). The bar plots represent the model accuracy, F1-score, and Cohen's κ for each cohort. Error bars indicate the 95% CIs while the center is always the computed value of each classification performance metric (specified by its respective axis labels). The slide-level performance is reported in Supplemental Table 6. The Swiss cohort was excluded from the analysis due to the absence of grade 3 rejections.



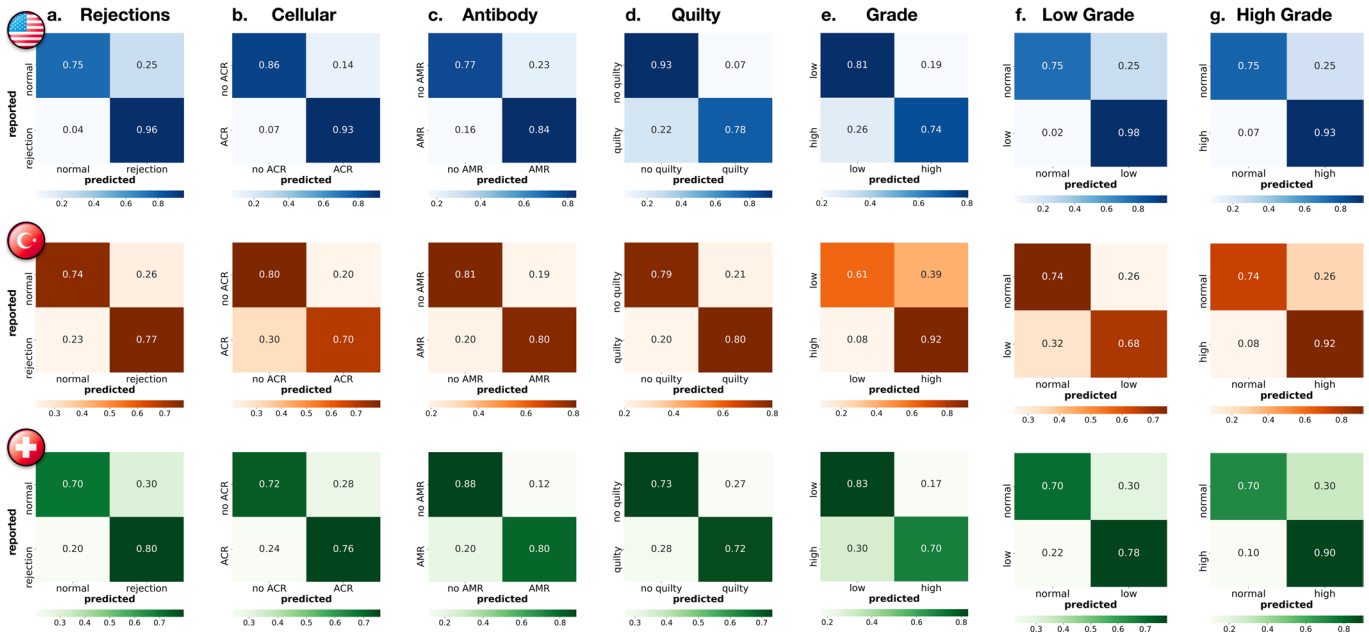
Extended Data Fig. 3 | Model performance at various magnifications. Model performance at different magnifications scales at **a**. slide-level and **b**. patient-level. Reported are AUC-ROC curves with 95% CI for 40×, 20× and 10× computed for the US test set (n = 995 WSIs, N = 336 patients). For the rejection detection tasks, the model typically performs better at higher magnification, while the grade predictions benefit from the increased context presented at lower magnifications. To account for the information from different scales, the detection of rejections and Quilty-B lesions is performed from the fusion of the model predictions from all available scales. In comparison, the rejection grade is determined from 10X magnification. **c**. Model performance during training and validation. Shown is cross-entropy loss for the multi-task model assessing the biopsy state and for the single-task model estimating the rejection grade. Reported is slide-level performance at 40× for the multi-task model, while the grading scores are measured at 10X magnification. The model with the lowest validation loss encountered during the training is used as the final model.



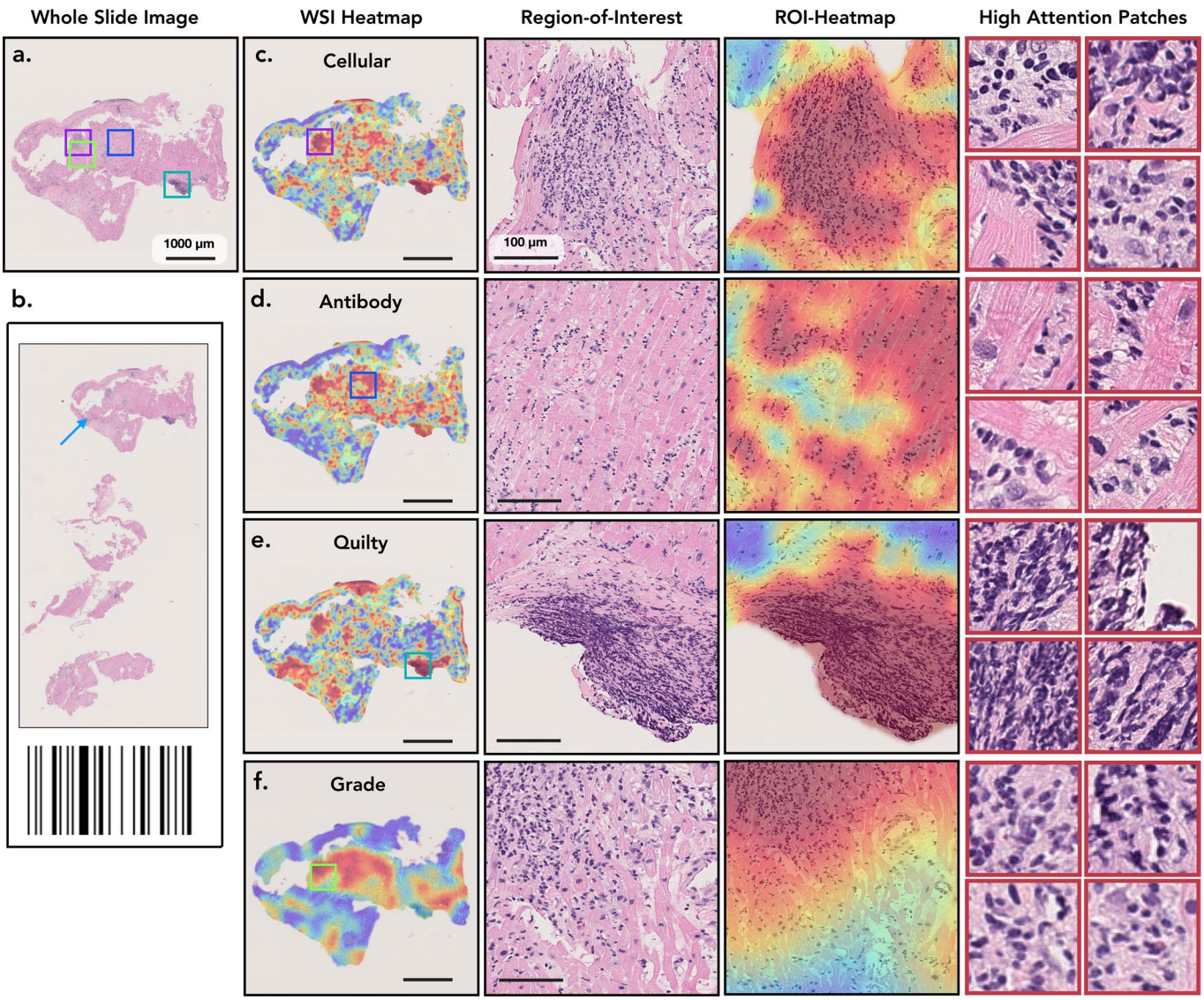
Extended Data Fig. 4 | Performance of the CRANE model at slide-level. The CRANE model was evaluated on the test set from the US (n = 995 WSIs, N = 336 patients) and two independent external cohorts from Turkey (n = 1,717, N = 585), and Switzerland (n = 123, N = 123). **a**. Receiver operating characteristic (ROC) curves for the multi-task classification of EMB and grading at the slide-level. The area under the ROC curve (AUC) scores are reported together with the 95% CIs. **b**. The bar plots reflect the model accuracy for each task. Error bars (marked by the black lines) indicate 95% CIs while the center is always the computed value for each cohort (specified by the respective axis labels). The results suggest the ability of the CRANE model to generalize across diverse populations, and different scanners and staining protocols, without any domain-specific adaptations. Clinical deployment might benefit from the model's fine-tuning with the local data and scanners.



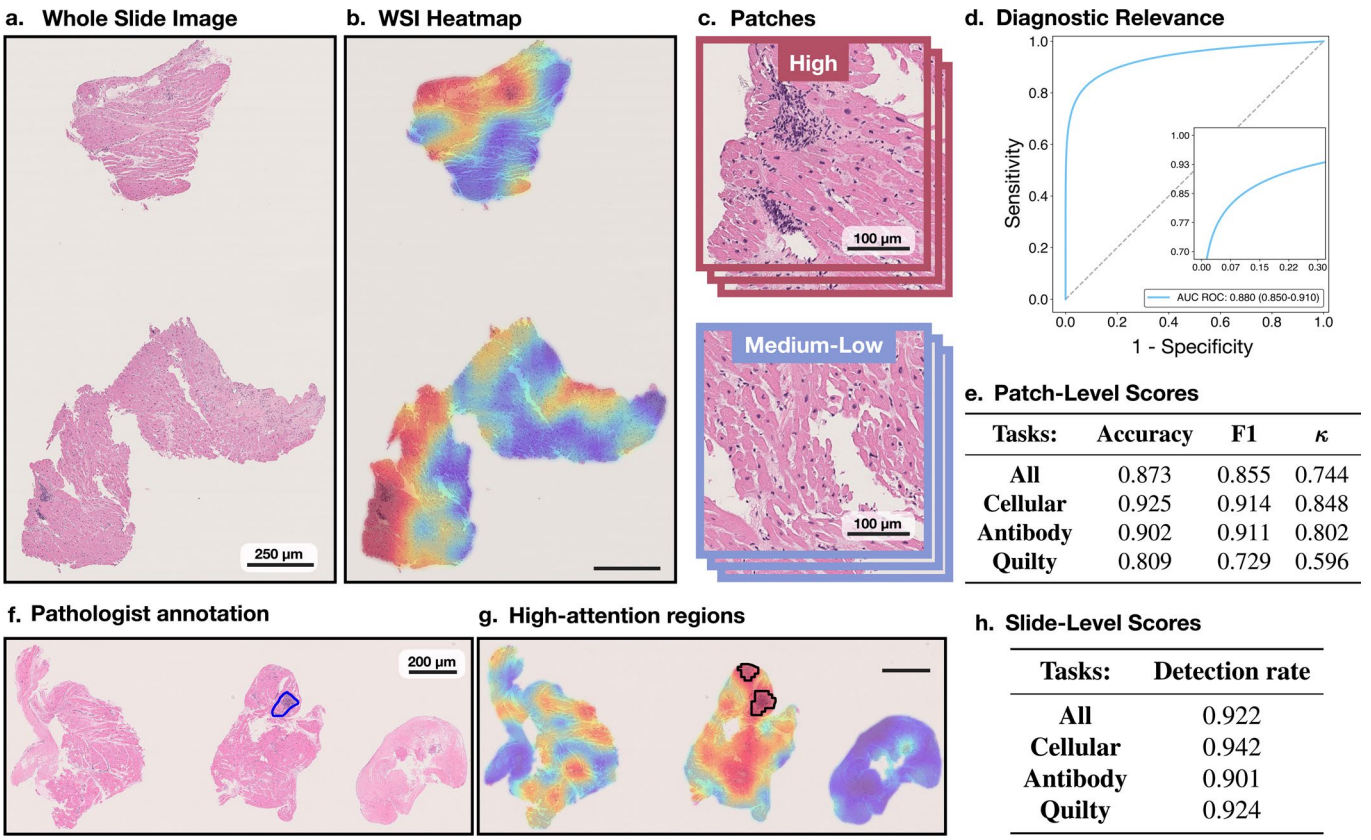
Extended Data Fig. 5 | Confidence of model's predictions. The model robustness can be measured through the confidence of the predictions. The models that suffer from overfitting usually reach high performance on the training dataset by memorizing the specifics of the training data rather than learning the task at hand. As a consequence, such models result in incorrect but highly confident predictions during the deployment. The bar plots show the fraction of model predictions achieved with high confidence, for both correctly (blue) and incorrectly (yellow) estimated patient cases. The fraction of highly confident correctly predicted samples is consistently higher than the fraction of confident incorrect predictions across all the tasks. These results indicate the robustness of the model predictions for all tasks.



Extended Data Fig. 6 | Patient-level performance for all prediction tasks. Reported are confusion matrices for **a**. rejection detection (including both ACR and AMR), detection of **b**. ACRs, **c**. AMRs, **d**. Quilty-B lesions, and **e**. discrimination between low (grade 1) and high (grade 2 + 3) rejections. To assess the model's ability to detect rejections of different grades, subplots **f**. shows the distinction between normal cases and low-grade rejections, while **g**. reports distinction between normal cases and high-grade rejections. In both external cohorts, the model reached higher performance for detecting the more clinically relevant high-grade rejections, whereas in the internal cohort the performance is comparable for both low and high-grade cases. The rows of the confusion matrices show the model predictions and the columns represent the diagnosis reported in the patient's records. The prediction cut-off for each task was computed from the validation set. For the clinical deployment, the cut-off can be modified and fine-tuned with the local data to meet the desirable false-negative rate. The performance is demonstrated on the US hold-out test set (N = 336 patients with 155 normal cases,181 rejections, 161 ACRs, 31 AMRs, 65 Quilty-B lesions, 113 low-grade, and 68 high-grade), Turkey (585 patients with 308 normal cases, 277 rejections, 271 ACRs, 16 AMRs, 74 Quilty-B lesions, 166 low-grade, and 111 high-grade) and Swiss (N = 123 patients with 54 normal cases, 69 rejections, 66 ACRs, 10 AMRs,18 Quilty-B lesions, 59 low-grade and 10 high-grade). Details on each cohort are reported in Supplemental Table 1.



Extended Data Fig. 7 | Analysis of case with concurrent cellular, antibody-mediated rejection, and Quilty-B lesions. a-b. The selected biopsy region and the corresponding H&E stained WSI. Attention heatmaps are computed for each task (**c,d,e**) and the grade (**f**). For the cellular task (**c**.), the high-attention regions correctly identified diffuse, multi-focal interstitial inflammatory infiltrate, predominantly comprised of lymphocytes, and associated myocyte injury. For the antibody heatmap (**d**.), the high-attention regions identified interstitial edema, endothelial swelling, and mild inflammation, consisting of lymphocytes and macrophages. For the Quilty-B heatmap (**e**.), the high-attention regions highlighted a focal, dense collection of lymphocytes within the endocardium, with mild crush artifact. For the grade (**f**.), the high-attention regions identified areas with diffuse, interstitial lymphocytic infiltrate with associated myocyte injury, corresponding to high grade cellular rejection. The high-attention regions for both types of rejection and Quilty-B lesions appear similar at the slide level at low power magnification, since all three tasks assign high-attention to regions with atypical myocardial tissue. However, at higher magnification, the highest attention in each task comes from regions with the task-specific morphology. The image patches with the highest attention scores from each task are shown in the last column. This example also illustrates the potential of CRANE to discriminate between ACR and similarly appearing Quilty-B lesions.



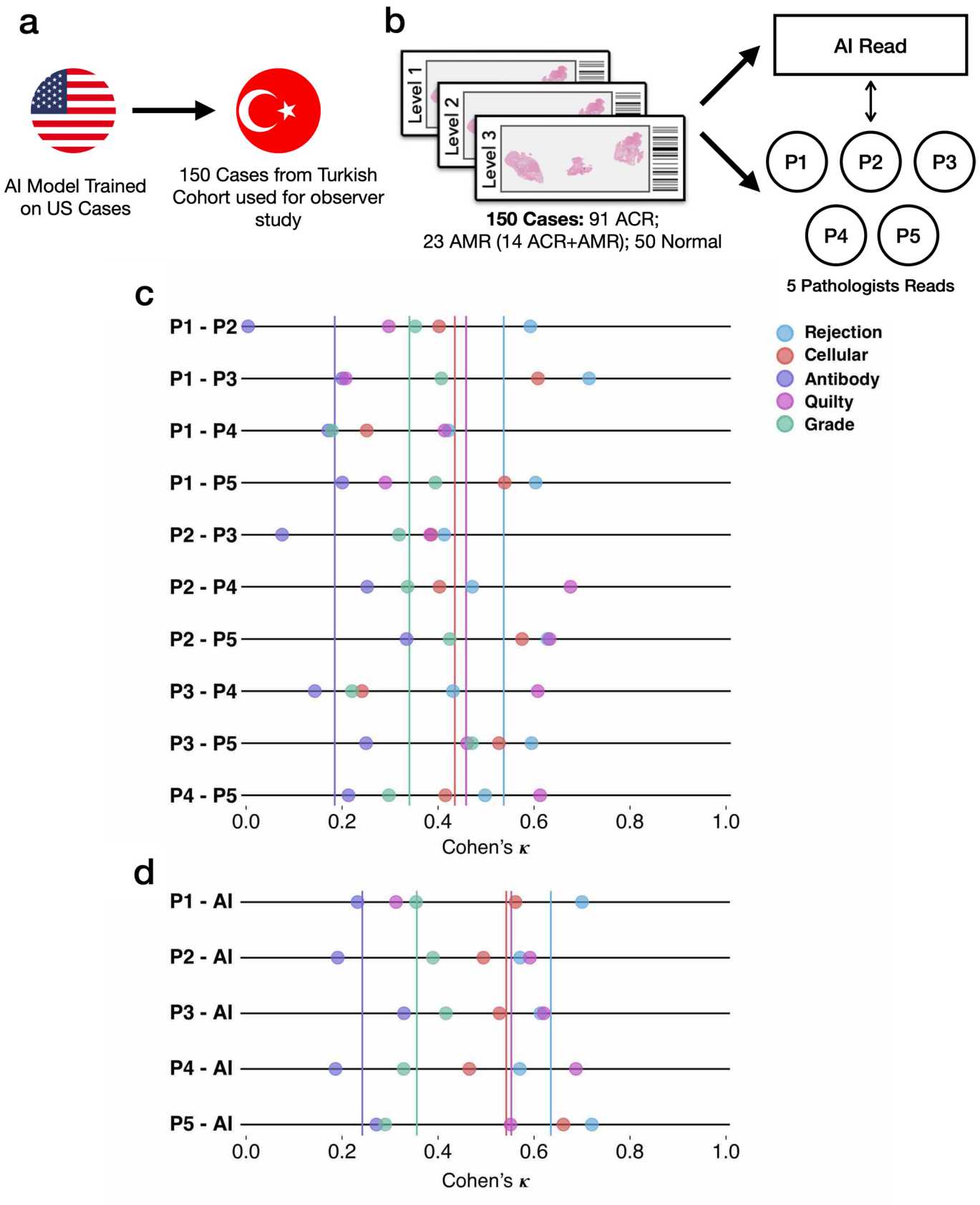
e. Patch-Level Scores

Tasks:	Accuracy	F1	κ
All	0.873	0.855	0.744
Cellular	0.925	0.914	0.848
Antibody	0.902	0.911	0.802
Quilty	0.809	0.729	0.596

h. Slide-Level Scores

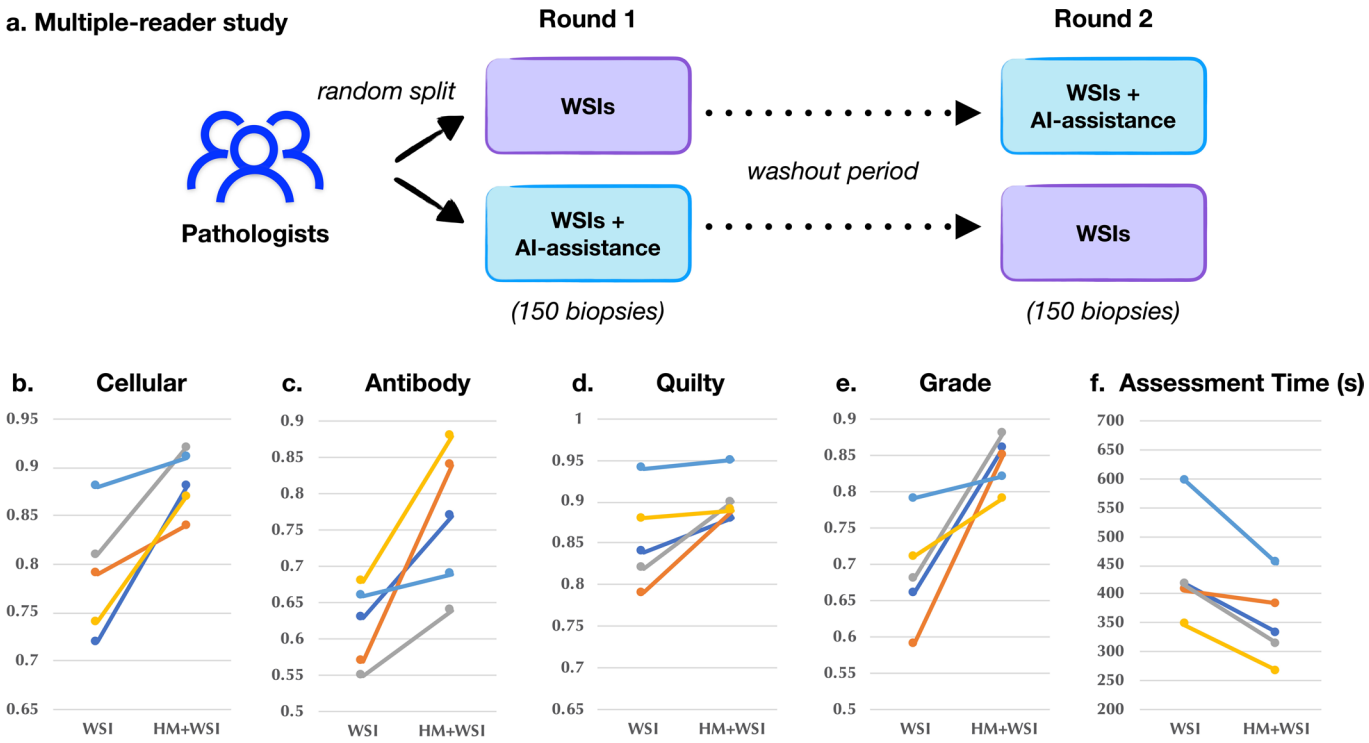
Tasks:	Detection rate
All	0.922
Cellular	0.942
Antibody	0.901
Quilty	0.924

Extended Data Fig. 8 | Quantitative assessment of attention heatmaps' interpretability. While the attention scores provide only relative importance of each biopsy regions for the model predictions, we attempted to quantify their relevance for diagnostic interpretability at patch- and slide-level. From the internal test set, we randomly selected 30 slides from each diagnosis and computed the attention heatmaps for each task (**a-b,f-g**).For the patch-level assessment, we selected 3 non-overlapping patches from the highest attention region in each slide. Since the regions with the lowest attention scores often include just a small fraction of tissue, we randomly selected 3 non-overlapping patches from the regions with medium-to-low attentions (i.e. attention scores<0.5). We randomly remove 5% of the patches to prevent pathologist from providing an equal amount of diagnoses, resulting in a total of 513 patches. A pathologist evaluated each patch as relevant or non-relevant for the given diagnosis. The pathologist's scores are compared against the model predictions of diagnostically relevant (high-attention) vs non-relevant (medium-to-low attention) patches. The subplot shows AUC-ROCscores across all patches, using the normalized attention scores as the probability estimates. The accuracy, F1-score, and Cohen's κ, computed for all patches and for the specific diagnoses, are reported in **e**.. These results suggest a high agreement between the model and pathologist's interpretation of diagnostically relevant regions. For the slide-level assessment, we compare concordance in the predictive regions used by the model and pathologists. A pathologist annotated in each slide the most relevant biopsy region(s) for the given diagnosis (**f**.). The regions with the top 10% highest attentions scores in each slide are used to determine the most relevant regions used by the model (**g**.). These are compared against the pathologist's annotations. The detection rate for all slides, and the individual diagnosis, are reported in **h**. Although the model did not use any pixel-level annotations during training these results imply relatively high concordance in the predictive regions used by the model and pathologist. It should be noted that the attention heatmaps are always normalized and not absolute, hence, the highest attended region is considered for the analysis similar to[17].



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Inter-observer variability analysis. The design of the reader study is depicted in **a-b**. The subplot **c**. shows the agreement between each pair of pathologists, while the agreement between the AI model and each pathologist is shown in **d**. The average agreement for each task is plotted as a vertical solid line. The analysis was performed on a random subset of 150 cases randomly selected from the Turkey test cohort: 91 ACR, 23 AMR cases (including 14 concurrent ACR and AMR cases) and 50 normal biopsies. The AI model was trained on the US cohort. For evaluation purposes, the pathologists assessed each case using the H&E slides only. It should be noted that the assessment presented here is based on Cohen's κ and is not the absolute agreement. Cohen's κ is a metric which runs between -1 and 1 and takes into account agreement by chance.



Extended Data Fig. 10 | AI-assisted biopsy assessment. An independent reader study was conducted to assess the potential of the CRANE to serve as an assisting diagnostic tool. Subplot **a**. illustrates the study design. A panel of five cardiac pathologists from an independent center was asked to assess 150 EMBs randomly selected from the Turkey cohort, the same set of slides as used for the assessment of interobserver variability presented in Extended Data Fig. 9. The pathologists were randomly split into two groups. In the first round, the readers from the first group used WSIs only, while the readers from the second group also received assistance from the CRANE in the form of attention heatmaps (HMs) plotted on the top of H&E slides. Following a washout period, the pathologists repeated the task. In the second round, the readers from the first group received WSIs and AI assistance, while the second group used WSIs only. Subplots **b-e**. report accuracy and assessment time (**f**.) of the readers without and with AI assistance marked as (WSI) and (HM + WSI), respectively. The ground truth labels were constructed based on the pathologists' consensus from the reader-study presented in Extended Data Fig. 9. The ability of the CRANE to mark diagnostically relevant regions has increased the accuracy of manual biopsy assessment for all tasks and all readers, as well as reduce the assessment time. These results support the feasibility of CRANE in reducing the interobserver variability and increasing the efficiency of manual biopsy reads.

nature portfolio

Corresponding author(s): Faisal Mahmood

Last updated by author(s): Dec 16 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All code associated with this project is available at https://github.com/mahmoodlab/CRANE
Data analysis	Analysis code was written in Python (3.7.5). The following Python libraries were used: hSpy (2.10.0), matplotlib (3.1.1), numpy (1.18.1), opencv-python (4.1.1), openslide-python (1.1.1), pandas (1.0.3), pillow (6.2.1), Pytorch (version 1.7.1), scikit-learn (0.22.1), scipy (1.3.1), tensorflow (1.14.0), tesnorboardx (1.9), torchvision (0.6), PytorchLightning (version 1.3.3) and timm (version 0.4.9). Additionally, R (version 3.6.1) and R library pROC library (version 1.16.2) were used. The WSIs were processed with CLAM software. The interactive demo website was developed using OpenSeadragon (version 2.4.2) and jQuery (version 3.6.0). The code used for this study has been made publicly available at: https://github.com/mahmoodlab/CRANE

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Please email all requests for academic use of raw and processed data to the corresponding author. Restrictions apply to the availability of the in-house and external data, which were used with institutional permission for the current study, and are thus not publicly available. Please email all requests for academic use of raw and

processed data to the corresponding author. All requests will be promptly evaluated based on institutional and departmental policies to determine whether the data requested is subject to intellectual property or patient privacy obligations. Data can only be shared for non-commercial academic purposes and will require a data user agreement. A subset of whole slide images used in the study can be accessed through our interactive demo available at http://crane.mahmoodlab.org. ImageNet data is available at https://image-net.org/

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculations were performed. For subtypes with abundant cases the deep learning models were trained while increasing the number of input slides until asymptotic improvement of model performance was achieved to suggest an appropriate sample size was obtained. For rare high grade rejections the sample size was limited by the number of cases available at the three medical centers. The number of cases from each cohort are described in the Methods section and demographics are available in Supplementary Table 1.
Data exclusions	Missing cases and cases with damaged slides were excluded. Cases with blurry scans were not excluded but rescanned.
Replication	Our training, test protocols and all code has been made publicly available for additional evaluation and reproducibility and may be accessed at https://github.com/mahmoodlab/CRANE
Randomization	We randomly split our dataset into 70% train, 10% validation, and 20% test splits on a patient level for the in house data. Randomization was conducted ensuring slides from the same case always remain in the same split. For the reader study, 150 cases were randomly selected from the Turkey cohort. The cases were presented in a computationally generated random order during the reader studies.
Blinding	For the reader studies all experts were blinded to the assessment from the other experts as well as to the group allocation during data collection and analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Population characteristics for the dataset are given in Supplementary Table 1.
Recruitment	No patient recruitment was necessary for using histology whole slide images retrospectively.
Ethics oversight	Institutional and ethical approval was obtained from USA and Turkey sites. The retrospective analysis of pathology slides was approved by the Mass General Brigham (MGB) IRB office under protocol 2020P000234 and Ege University, Turkey under protocol 20-11T/61. The samples from Switzerland were exempt from the need for IRB approval according to the Swiss Human Research Act (HFG 810.30, 30. September 2011). At all three sites, informed consent was waived for retrospective analysis of historical pathology slides and corresponding pathology reports.

Note that full information on the approval of the study protocol must also be provided in the manuscript.