Prediction of heart transplant rejection from routine pathology slides with self-supervised Deep Learning

Tobias Paul Seraphin (1,3*), Mark Luedde (2,*), Christoph Roderburg (1,*), Marko van Treeck (3), Pascal Scheider (4), Roman D. Buelow (4), Peter Boor (4), Sven H. Loosen (1), Zdenek Provaznik (5), Daniel Mendelsohn (6), Filip Berisha (7,8), Christina Magnussen (7,8), Dirk Westermann (7,8), Tom Luedde (1), Christoph Brochhausen (6,*), Samuel Sossalla (9,10,11,*), Jakob Nikolas Kather (3,12,13,14,*)

* equal contribution

(1) Department of Gastroenterology, Hepatology and Infectious Diseases, University Hospital Duesseldorf, Medical Faculty at Heinrich-Heine-University, Dusseldorf, Germany.
(2) Department of Cardiology and Angiology, Christian-Albrechts-University of Kiel, Kiel, Germany.
(3) Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany.
(4) Institute of Pathology, RWTH Aachen University Hospital, Aachen, Germany.
(5) Department of cardiothoracic surgery, University Medical Center Regensburg, Regensburg, Germany
(6) Institute of Pathology, University of Regensburg, Regensburg, Germany.
(7) Department of Cardiology, University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Hospital Eppendorf, Hamburg, Germany.
(8) German Center for Cardiovascular Research (DZHK), Partner Site Hamburg/Kiel/Lübeck, Germany.
(9) Clinic for Cardiology and Pneumology, Georg-August University Göttingen
(10) DZHK (German Center of Cardiovascular Research), Partner Site Göttingen, Göttingen, Germany
(11) Department of Internal Medicine II, University Medical Center Regensburg, Regensburg, Germany.
(12) Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany
(13) Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, United Kingdom
(14) Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany.

Correspondence to:
Jakob Nikolas Kather, MD, MSc
Professor of Clinical Artificial Intelligence
Else Kroener Fresenius Center for Digital Health
Technical University Dresden
Fetscherstrasse 74, 01307 Dresden, Germany
jakob-nikolas.kather@alumni.dkfz.de

Abstract

Background and Aims

One of the most important complications of heart transplantation is organ rejection, which is diagnosed on endomyocardial biopsies by pathologists. Computer-based systems could assist in the diagnostic process and potentially improve reproducibility. Here, we evaluated the feasibility of using deep learning in predicting the degree of cellular rejection from pathology slides as defined by the International Society for Heart and Lung Transplantation (ISHLT) grading system.

Methods

We collected 1079 histopathology slides from 325 patients from three transplant centers in Germany. We trained an attention-based deep neural network to predict rejection in the primary cohort and evaluated its performance using cross validation and by deploying it to three cohorts.

Results

For binary prediction (rejection yes/no) the mean Area Under the Receiver Operating Curve (AUROC) was 0.849 in the cross-validated experiment and 0.734, 0.729 and 0.716 in external validation cohorts. For a prediction of the ISHLT grade (0R, 1R, 2/3R), AUROCs were 0.835, 0.633 and 0.905 in the cross-validated experiment and 0.764, 0.597, 0.913, and 0.631, 0.633, 0.682, and 0.722, 0.601, 0.805 in the validation cohorts, respectively. The predictions of the AI model were interpretable by human experts and highlighted plausible morphological patterns.

Conclusions

We conclude that artificial intelligence can detect patterns of cellular transplant rejection in routine pathology, even when trained on small cohorts.

Introduction

In patients with end-stage heart failure, organ transplantation constitutes the desired curative treatment concept [1]. This has been made possible in recent decades, in particular, by the advent of new immunosuppressive drugs, which can ensure long-lasting organ preservation. However, organ rejection by the host immune system remains one of the major complications in these patients [2]. Despite the increasing importance of noninvasive methods in the detection of graft rejection, endomyocardial biopsy remains the gold standard for detecting rejection, especially in the first year after transplantation [3,4]. The pathological assessment of such specimens is reserved for highly specialized pathologists and has massive clinical consequences. In 1990 the International Society for Heart and Lung Transplantation (ISHLT) published a guideline for histopathologic diagnosis of acute cellular rejection to standardize this assessment, which has been revised in 2004 [5]. Nevertheless, the purely subjective assessment of pathological sections has certain disadvantages, such as the dependency on appropriately trained experts, as well as remaining inter- and intra-observer variability [6]. In addition, endomyocardial biopsies are obtained either as routine surveillance protocol diagnostics or as a diagnostic investigation in patients with allograft dysfunction and clinically suspected rejection.

Computer-based image analysis programs can potentially support pathology experts in performing diagnostics. In several histopathological applications, it could be shown that such computer-based image analysis programs can show a high level of concordance with human observers, and in some cases the combination with the human experts can improve the consistency of the findings.[7] In particular, the technology of artificial neural networks has brought very good results in many clinically relevant prediction tasks in recent years [8,9]. A recent extension of this technology is the so-called attention-based multiple instance learning [10], in which the artificial neural network can learn which areas of the whole slide image are more relevant than other areas [11,12].

In contrast to solid tumors, in which many studies have examined computer-based prediction of clinically relevant biomarkers in the last three years [9], there are only comparatively few studies in the context of transplantation medicine. Precedent cases exist in the prediction of organ rejection after kidney transplantation [13], as well as applications of simple, handcrafted feature based image analysis methods to cardiac biopsies after transplantation [14,15]. A recent study by Lipkova et al. used the Deep Learning pipeline “CRANE” to predict cardiac allograft transplantation, yielding a very high and clinical-grade performance [16,17].

However, several open questions remain regarding the data requirements to train such systems, as Lipkova et al. trained their system on thousands of patient samples, but this large number of samples is rarely available. Additional questions remain open regarding the generalizability of such systems, and the biological interpretability which can be drawn from their predictions. Finally, new technical approaches such as self-supervised learning (SSL) to pre-train pathology Deep Learning models could yield an improved performance [18], but this has not yet been evaluated in prediction of cardiac allograft rejection.

In the present study, we collected four cohorts from three hospitals of cardiac transplant patients undergoing cardiac biopsy routinely and based on clinically relevant changes. We trained our own SSL-attention-based Deep Learning pipeline as well as CRANE on these data and evaluated the predictive performance regarding the presence of cellular transplant rejection.

Materials and Methods

Patient cohorts and experimental design

In this study we included four case series ("patient cohorts") from three different medical centers in Germany. The first cohort was obtained from the pathological archive of the University Hospital Regensburg and contained 393 pathological sections from 107 patients from the period 2016 to 2018. The second cohort also originated from the pathological archive of the University Hospital Regensburg and contained 356 pathological sections from 95 patients from the period 2019 to 2021. The third cohort was obtained from the pathological archive of the University Medical Center Hamburg-Eppendorf. This cohort contained 189 pathological sections from 86 patients from the period 2019 to 2021. The fourth cohort was obtained from the pathological archive of the University Hospital Aachen containing 141 pathological sections from 37 patients from the period 1999 to 2014. Cohorts were consecutive retrospective case series. We pragmatically aimed to maximize the sample size of training and testing cohorts. The ground truth was obtained by two expert pathologists during routine work-up at each participating center, grading the degree of rejection in consensus, following the 2004 revision of the ISHLT grading system.[5] All patient samples without information on ISHLT grading were not eligible for inclusion. A detailed presentation of the clinical characteristics of all patients in the corresponding cohorts can be found in **Table 1**. We used two categorizations of the ISHLT 2004 grading system as our prediction target. The first is a binarized target ("ISHLT 2004 rejection "yes/no"), summarizing slides with ISHLT 0R on the one hand (class "no") and all signs of rejection on the other hand (ISHLT 1R, 2R, 3R; class "yes"). For the second target ("ISHLT 2004 rejection grade") we aimed for a more granular classification splitting the second class giving three classes comprising ISHLT 0R, ISHLT 1R and ISHLT 2R & 3R. We combined the higher order rejection due to shortage of ISHLT 3R cases in the training set (**Table 1**). Our study adheres to the STARD guidelines (**Suppl. Table 1**). [19]

Sample processing and image preprocessing

Routine tissue sections were obtained from the pathology archives at the above-mentioned institutions. All slides were stained with hematoxylin and eosin (H&E) according to standard clinical protocols at each center. Pen marks were removed from the slides of the training cohort. All images were digitized at 40x magnification with an Aperio AT2 Slide scanner (Aperio, Leica Camera AG, Wetzlar, Germany) centrally at the University Hospital Düsseldorf (**Figure 1a**). All images were available in ScanScope Virtual Slide (SVS) format and were tessellated in tissue patches of 512x512 pixels size using https://github.com/KatherLab/preprocessing-ng according to the "The Aachen Protocol for Deep Learning Histopathology: A hands-on guide for data preprocessing" (**Figure 1b**) [20].

Deep Learning workflow

For all Deep Learning experiments, we used our in-house pipeline "Marugoto", which is publicly available at https://github.com/KatherLab/marugoto and has been previously used for analysis of images obtained from cancer tissue [21]. In this approach, each image tile was translated into a 2048-dimensional feature vector by a pre-trained histology-specific encoder RetCCL (https://github.com/Xiyue-Wang/RetCCL). We used attention-based multiple instance learning [22], in which all feature vectors obtained from all tiles from one whole slide image constitute a "bag" which is processed by the neural network (**Figure 1b**). The multiple instance learning network is structured as follows: The feature vectors of each of the bag's tiles are first projected into a length 256 feature space using a fully connected layer. Based on these, an attention module consisting of two fully

connected layers calculates an attention score for each of the tiles. All of a bag's attention scores are then normalized using softmax. We then calculate a bag-level feature vector by taking the sum of the tiles' feature projections weighted by their respective attention scores. The final classification is then done with an additional fully connected layer (**Figure 1c**). During training, we limited our bag size to a maximum of 512 tiles from each slide, resampled in each epoch (median number of tiles = 403, interquartile range = 468). We stopped the training of our model if no reduction in the validation loss was present for 16 following epochs while training for a maximum of 32 epochs. For deployment, we used all of the slides' tiles. We compared our approach to CRANE as presented by Lipkova et al. [16]. To do so, we followed the workflow of the CRANE study, preprocessing the slides with the CLAM repository and performed 10-fold Monte-Carlo cross validation on our training cohort, deploying the best performing model on our test cohorts [23].

Experimental design and hardware

We pre-specified the following experimental design. First we trained and evaluated our system in the first cohort via three-fold cross validation and repeated this experiment five times. Subsequently, we evaluated the performance of the best performing model on the second, third and fourth cohort (**Figure 1d**). All ground truth labels were available on the level of slides. All statistics were calculated on the level of slides. The primary evaluation metric was the Area Under the Receiver Operating Curve (AUROC). We calculated the mean performance as the mean of all AUROCs from all folds of all repetitions, together with the 95% confidence interval (95% CI) calculated by assuming a normalized distribution of AUROCs and using its standard error of the mean to identify the boundaries. For the multiclass prediction we used micro-averaging to obtain an overall AUROC of the experiments. We calculated p-values for each class in each experiment using a two sided t-test and averaged these values over folds and repetitions of the experiments. For visualization approaches, we deployed the best performing model on the test cohorts. All experiments were run on local desktop workstations with Nvidia RTX Quadro 8000 graphics processing units (GPUs).

Visualization and explainability

We plotted three tiles for the four slides of each validation cohort giving the highest bag label scores for the binarized prediction of (true) rejection when deploying the best performing model. Additionally we generated Grad-CAM images for these tiles to get a better understanding of the models attention.[24] To gain further insight into our model's decisions, we generated heatmaps showing the attention, as well as the attention multiplied by the prediction scores.

Code availability

All source codes for preprocessing are available at https://github.com/KatherLab/preprocessing-ng. All source codes for Deep Learning are available at https://github.com/KatherLab/marugoto.

Results

Deep learning can predict rejection and rejection grade from pathology images

We trained an attention-based multiple-instance deep learning algorithm on bags of features, extracted from patches of whole slide images. In the cross-validated experiment carried out on cohort 1, we found a mean AUROC of 0.849 (95% CI 0.822 - 0.877) for binary prediction (rejection yes/no) (**Figure 2a,** see **Suppl. Table 2** for individual results). The best fold's AUROC was 0.910 with a p-

value of <0.001. For prediction of the ISHLT grades 0R, 1R and 2/3R, the mean AUROCs were 0.835 (95% CI 0.807 - 0.862), 0.633 (95% CI 0.582 - 0.684) and 0.905 (95% CI 0.874 - 0.937), respectively (**Figure 2b,** see **Suppl. Table 3** for individual results). The micro-averaged AUROC for this task was 0.814 (95% CI 0.773 - 0.854). The best-fold's AUROCs for this task were 0.890, 0.808 and 0.968, respectively, with a p-value <0.001 and a micro-averaged AUROC of 0.885. These results show the capacity of our network to predict rejection and rejection grade directly from histopathology images.

Deep learning classifiers generalize to held-out and external patient cohorts

To further validate the performance of our network, we deployed the best performing model for each target on three validation cohorts. The validation experiments for cohort 2 yielded an AUROC of 0.734 (p-value <0.001) for binary prediction (rejection yes/no) (**Figure 2c**). For prediction of the ISHLT grades 0R, 1R and 2/3R, the AUROCs were 0.764 (p-value <0.001), 0.597 (p-value 0.099) and 0.913 (p-value <0.001) (**Suppl. Figure 1a**), respectively. The micro-averaged AUROC was 0.731 (p-value 0.021). For external validation on cohort 3 we obtained an AUROC of 0.729 (p-value of <0.001) (**Figure 2d**). For prediction of the ISHLT grades 0R, 1R and 2/3R, the AUROCs were 0.631 (p-value <0.013), 0.595 (p-value 0.048) and 0.682 (p-value 0.082), respectively (**Suppl. Figure 1b**). The micro-averaged AUROC was 0.659 (p-value 0.025). The external validation on cohort 4 yielded an AUROC of 0.716 (p-value <0.001) on the binary task (rejection yes/no) (**Figure 2e**). For prediction of the ISHLT grades 0R, 1R and 2/3R, the AUROCs were 0.722 (p-value <0.001), 0.601 (p-value 0.247) and 0.805 (p-value <0.001), respectively (**Suppl. Figure 1c**). The micro-averaged AUROC was 0.737 (p-value 0.042). Our findings show that our models are in principle generalizable to external patient cohorts.

Comparison of the deep learning classifier with CRANE

We compared our method to CRANE, the current state of the art in rejection prediction of heart transplant tissue slides [16]. In the training cohort, the cross-validated mean AUROC of the CRANE models for the binarized target (rejection yes/no) was 0.776 (95% CI 0.717 - 0.835) (**Figure 2f,** see **Suppl. Table 4** for individual results), lower than the performance obtained by our attention-MIL pipeline (0.849). The best performing CRANE model yielded an AUROC of 0.882, which was again slightly lower than the performance achieved by our in-house attention-MIL pipeline (0.910). When deploying the CRANE model to our test cohorts we received AUROCs of 0.831, 0.616, and 0.483 for cohorts 2, 3 and 4, respectively (**Figure 2g**), overall underperforming compared to our SSL-attention model (which reached 0.734, 0.729 and 0.716, respectively). In summary, our findings show that SSL-attention-MIL outperforms CRANE.

Attention-based predictions are explainable

To make the model's prediction explainable and to identify reasons for failure cases, we performed a reverse engineering task to see the spatial distribution of the network's attention layer for the most confident true classification of binary prediction. First of all our attention maps show that our model is concentrating only on tissue regions and not on the background or artifacts (See **Figure 3c** and **3g**). This means that the presence of such artifacts (e.g., pen marks) in the test set is not problematic, and that only a simple quality control algorithm might be sufficient for clinical implementation. Analyzing whole slide attention and prediction maps on a higher resolution, we found that our model's focus lies mainly on regions with a high lymphocyte density. Yet it seems to focus more on the interface of lymphocyte aggregations with the neighboring myocardium than on these dense regions themselves (see **Figure 3**). We also found evidence that our model apparently was confused by the presence of a Quilty lesion [25], which was observed in a misclassified patient (**Suppl. Figure 2**).

When analyzing the top tiles and the corresponding Grad-CAM images of the external validation cohorts, it seems that the model is concentrating on lymphocytes, confirming the findings made in heatmaps at another spatial scale (**Figure 4**). These findings show that despite being trained on only a few hundred patients, the model has learned clinically relevant morphological patterns from whole slide images.

Discussion

Heart transplantation remains the gold standard therapy for end-stage heart failure [26]. Due to this pronounced shortage of donor organs, there is not only a need for risk adjustment tools to optimize recipient selection [27,28]. In addition, a particularly good risk stratification and early adjustment of immunosuppression therapy is necessary in organ recipients because the possibility of re-transplantation is very limited. New diagnostic methods based on artificial intelligence (AI) could change and improve medical decision making in transplantation medicine in the future.[17] A potential key benefit would be to reduce diagnostic uncertainty, and hence reduce the need for frequent re-biopsies in the first year after transplantation, which represents a burden for healthcare systems and patients alike.

In the present study, we trained an AI method to evaluate the recognition and grading of cardiac transplantation using routine biopsies. We found high performance in the training set (by cross-validation). When deploying our model at the external validation cohorts, we found a stable, but moderate performance. A few other studies have addressed similar problems in recent years. Peyster et al. used handcrafted features to grade cellular rejection reporting good performance, already in 2021 [15]. Most prominently, Lipkova et al. presented the CRANE method, which yielded very high AUROCs in their study, after being trained on thousands of patients [16]. Lipkova et al. report an external validation AUROC of around 0.83, which is better than the AUROCs of around 0.72 which we report in the validation cohorts [16]. However, our training dataset comprised 10 times fewer patients, and in a head-to-head comparison of CRANE and our SSL-attention, our method outperforms CRANE, pointing to a higher data efficiency. Our findings are in line with other recent studies showing the usefulness of pre-training feature extractors with SSL, boosting classification performance in computational pathology [18]. Our classifier also outperforms other studies which date back to the year 2017, when Tong et al. constructed a shallow neural network based on handcrafted features derived from 43 WSIs (Children's Healthcare of Atlanta cohort). This dataset has been used several times afterwards improving the performance of the cross validated model while adopting newer methodology but remains limited due to the very small dataset size [14,29–31].

A fundamental limitation affecting all published studies is the limitation of the gold standard. The ISHLT classification itself is an imperfect predictor of clinical outcome, and future studies should train AI models directly on outcome data to overcome these limitations. This is further supported by the observation that detection and grading of heart transplant rejection can suffer from a suboptimal concordance among pathologists in the assignment of ISHLT 2004 grading of 71%, with most agreement coming from the class 0R [32]. While our study does not directly show that AI can improve objectivity and concordance, future studies should investigate the performance of pathologists who are guided by the AI model, especially non-expert pathologists.

In summary, our study is a proof of concept that shows the potential of AI systems in transplantation medicine. In particular, our study sets a new technical state of the art, which however requires validation in larger cohorts. On the other hand, our study is also a reminder that larger training cohorts of a few thousand patients are probably required for clinical-grade AI biomarkers [33,34].

Future studies should compare our technical approach on larger cohorts, which could be efficiently assembled with federated or Swarm Learning [35,36]. Our study adds to the growing evidence of AI models being capable of recognizing heart transplant rejection which might in the future help pathologists with prescreening slides or standardize grading across different centers. We also believe that further development of our approach harbors the potential to ultimately reduce costs and time in this sector of the healthcare system.

Additional information

Author contributions

TPS, ML, CR, CB, SS and JNK conceived the idea for the study. TPS and JNK performed the experiments. MVT developed the software. PS, RDB, PB, ZP, DM, FB, CM, DW, CB and SS contributed tissue samples. TPS and JNK wrote the initial draft of the manuscript. All authors contributed to the interpretation of the results, the editing of the final manuscript and gave approval for submission.

Declarations

Competing interests

JNK declares consulting services for Owkin, France and Panakeia, UK as well as reimbursement for scientific talks by MSD and Eisai. DW declares consulting services and honorary talks for Abiomed, AstraZeneca, Bayer, Berlin-Chemie, Novartis, Medtronic. CM declares honorary talks for AstraZeneca, Novartis, Heinen&Loewenstein, Boehringer Ingelheim/Lilly, Bayer, Pfizer, Sanofi, Aventis, Apontis, Abbott and meeting support from AstraZeneca, Novartis, Boehringer Ingelheim/Lilly. TL declares consulting fees from AstraZeneca, BMS, EISAI, Incyte, MSD, Roche, HepaRegeniX and honorary talks and travel support from Abbvie and Gilead. The other authors do not have anything to disclose.

Ethics approval

This study was carried out in accordance with the Declaration of Helsinki. This study is a retrospective analysis of digital images of anonymized archival tissue samples from three patient cohorts. Collection and anonymization of patients in all cohorts took place in each contributing centre. For the Regensburg cohorts, ethical approval was granted by the ethical Review board of the University Regensburg (ID: 21-2620-104). For the Hamburg and Aachen cohort, ethical approval of this retrospective investigation was not legally required due to local regulations (Berufsordnung fuer Aerzte). Nevertheless, the retrospective analysis of samples from Aachen and other collaborating centers was assessed by and approved by the Ethics commission of the Medical Faculty of RWTH Aachen University (EK315/19), confirming that no specific patient consent is required for this retrospective study of anonymized tissue samples.

Funding

JNK and TL are supported by the German Federal Ministry of Health (DEEP LIVER, ZMVI1-2520DAT111) and JNK is supported by the Max-Eder-Programme of the German Cancer Aid (grant #70113864). TL is supported by the European Research Council (ERC; Consolidator Grant No 771083). SS is funded by the German Research Foundation (DFG) through the research grant SO 1223/4-1. PB is supported by the German Research Foundation (DFG; Project-IDs 322900939, 454024652, 445703531), the European Research Council (ERC; Consolidator Grant No 101001791), the federal Ministry of Health (DEEP LIVER, ZMVI1-2520DAT111) and the Federal Ministry of Economic Affairs and Energy (EMPAIA, No. 01MK2002A). ML is supported by the German Foundation for the chronically ill (III. Grant). CM is funded by the German Center for Cardiovascular Research, *Deutsche Stiftung für Herzforschung* und *Rolf M. Schwiete Stiftung*. This research project is supported by the START-Program of the Faculty of Medicine of the RWTH Aachen University (148/21 to RDB).

References

1. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J. 2021;42: 3599–3726.

2. Lund LH, Khush KK, Cherikh WS, Goldfarb S, Kucheryavaya AY, Levvey BJ, 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. 2017;36: 1037–1046.

3. Ruiz-Ortiz M, Rodriguez-Diego S, Delgado M, Kim J, Weinsaft JW, Ortega R, et al. Myocardial deformation and acute cellular rejection after heart transplantation: Impact of inter-vendor variability in diagnostic effectiveness. Echocardiography. 2019;36: 2185–2194.

4. van Heeswijk RB, Piccini D, Tozzi P, Rotman S, Meyer P, Schwitter J, et al. Three-dimensional self-navigated T2 mapping for the detection of acute cellular rejection after orthotopic heart transplantation. Transplant Direct. 2017;3: e149.

5. Stewart S, Winters GL, Fishbein MC, Tazelaar HD, Kobashigawa J, Abrams J, et al. Revision of the 1990 working formulation for the standardization of nomenclature in the diagnosis of heart rejection. J Heart Lung Transplant. 2005;24: 1710–1720.

6. Angelini A, Andersen CB, Bartoloni G, Black F, Bishop P, Doran H, 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. 2011;30: 1214–1220.

7. Tizhoosh HR, Diamandis P, Campbell CJV, Safarpoor A, Kalra S, Maleki D, et al. Searching Images for Consensus: Can AI Remove Observer Variability in Pathology? Am J Pathol. 2021;191: 1702–1708.

8. Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN. Deep learning in cancer pathology: a new generation of clinical biomarkers. British Journal of Cancer. 2020. doi:10.1038/s41416-020-01122-x

9. Cifci D, Foersch S, Kather JN. Artificial intelligence to identify genetic alterations in conventional histopathology. J Pathol. 2022. doi:10.1002/path.5898

10. Ilse M, Tomczak JM, Welling M. Attention-based Deep Multiple Instance Learning. arXiv [cs.LG]. 2018. Available: http://arxiv.org/abs/1802.04712

11. Shmatko A, Ghaffari Laleh N, Gerstung M, Kather JN. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. Nat Cancer. 2022;3: 1026–1038.

12. Ghaffari Laleh N, Muti HS, Loeffler CML, Echle A, Saldanha OL, Mahmood F, et al. Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. Med Image Anal. 2022;79: 102474.

13. Kers J, Bülow RD, Klinkhammer BM, Breimer GE, Fontana F, Abiola AA, et al. Deep learning-based classification of kidney transplant pathology: a retrospective, multicentre, proof-of-concept study. Lancet Digit Health. 2022;4: e18–e26.

14. Tong L, Hoffman R, Deshpande SR, Wang MD. Predicting heart rejection using histopathological whole-slide imaging and deep neural network with dropout. 2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI). 2017. pp. 1–4.

15. Peyster EG, Arabyarmohammadi S, Janowczyk A, Azarianpour-Esfahani S, Sekulic M, Cassol

C, et al. An automated computational image analysis pipeline for histological grading of cardiac allograft rejection. Eur Heart J. 2021;42: 2356–2369.

16. Lipkova J, Chen TY, Lu MY, Chen RJ, Shady M, Williams M, et al. Deep learning-enabled assessment of cardiac allograft rejection from endomyocardial biopsies. Nat Med. 2022;28: 575–582.

17. Mahmood F, Topol EJ. Digitising heart transplant rejection. Lancet. 2022;400: 17.

18. Schirris Y, Gavves E, Nederlof I, Horlings HM, Teuwen J. DeepSMILE: Contrastive self-supervised pre-training benefits MSI and HRD classification directly from H&E whole-slide images in colorectal and breast cancer. Med Image Anal. 2022;79: 102464.

19. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351: h5527.

20. Muti HS, Loeffler C, Echle A, Heij LR, Buelow RD, Krause J, et al. The Aachen Protocol for Deep Learning Histopathology: A hands-on guide for data preprocessing. Zenodo; 2020. doi:10.5281/ZENODO.3694994

21. Saldanha OL, Loeffler CML, Niehues JM, van Treeck M, Seraphin TP, Hewitt KJ, et al. Self-supervised deep learning for pan-cancer mutation prediction from histopathology. bioRxiv. 2022. p. 2022.09.15.507455. doi:10.1101/2022.09.15.507455

22. Ilse M, Tomczak J, Welling M. Attention-based Deep Multiple Instance Learning. In: Dy J, Krause A, editors. Proceedings of the 35th International Conference on Machine Learning. PMLR; 10--15 Jul 2018. pp. 2127–2136.

23. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng. 2021;5: 555–570.

24. Selvaraju, Cogswell, Das. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proc Estonian Acad Sci Biol Ecol. Available: http://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html

25. Forbes RD, Rowan RA, Billingham ME. Endocardial infiltrates in human heart transplants: a serial biopsy analysis comparing four immunosuppression protocols. Hum Pathol. 1990;21: 850–855.

26. McDonagh TA, Metra M, Adamo M. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the Task Force for the diagnosis and treatment of acute and …. Eur Heart J. 2021. Available: https://academic.oup.com/eurheartj/article-abstract/42/36/3599/6358045

27. Schramm R, Zittermann A, Fuchs U, Fleischhauer J, Costard-Jäckle A, Ruiz-Cano M, et al. Donor-recipient risk assessment tools in heart transplant recipients: the Bad Oeynhausen experience. ESC Heart Fail. 2021;8: 4843–4851.

28. Sunavsky J, Fujita B, Ensminger S, Börgermann J, Morshuis M, Fuchs U, et al. Predictors of failure after high urgent listing for a heart transplant. Interact Cardiovasc Thorac Surg. 2018;27: 950–957.

29. Dooley AE, Tong L, Deshpande SR, Wang MD. Prediction of Heart Transplant Rejection Using Histopathological Whole-Slide Imaging. IEEE EMBS Int Conf Biomed Health Inform. 2018;2018. doi:10.1109/bhi.2018.8333416

30. Zhu Y, Wang MD, Tong L, Deshpande SR. Improved Prediction on Heart Transplant Rejection Using Convolutional Autoencoder and Multiple Instance Learning on Whole-Slide Imaging. IEEE EMBS Int Conf Biomed Health Inform. 2019;2019. doi:10.1109/bhi.2019.8834632

31. Giuste F, Venkatesan M, Zhao C, Tong L, Zhu Y, Deshpande SR, et al. Automated Classification of Acute Rejection from Endomyocardial Biopsies. Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. New York, NY, USA: Association for Computing Machinery; 2020. pp. 1–9.

32. Crespo-Leiro MG, Zuckermann A, Bara C, Mohacsi P, Schulz U, Boyle A, et al. Concordance among pathologists in the second Cardiac Allograft Rejection Gene Expression Observational Study (CARGO II). Transplantation. 2012;94: 1172–1177.

33. Echle A, Grabsch HI, Quirke P, van den Brandt PA, West NP, Hutchins GGA, et al. Clinical-Grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning. Gastroenterology. 2020;159: 1406–1416.e11.

34. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019;25: 1301–1309.

35. Lu MY, Chen RJ, Kong D, Lipkova J, Singh R, Williamson DFK, et al. Federated learning for computational pathology on gigapixel whole slide images. Med Image Anal. 2022;76: 102298.

36. Saldanha OL, Quirke P, West NP, James JA, Loughrey MB, Grabsch HI, et al. Swarm learning for decentralized artificial intelligence in cancer histopathology. 2021. Available: https://europepmc.org/article/ppr/ppr423063

Tables

	Cohort 1	**Cohort 2**	**Cohort 3**	**Cohort 4**
Contributing centre	Regensburg	Regensburg	Hamburg	Aachen
Use in this study	Train	Test	Test	Test
N patients	107	95	86	37
N slides	393	356	189	141
Recruitment years	2016 - 2018	2019 - 2021	2019 - 2021	1999 - 2014
Age in years (median, IQR)	N/A	61 (10)	52 (17)	55 (12)
Gender (F:M)	N/A	N/A	70:121	47:94
ISHLT rejection **no** **yes**	 312 81	 271 85	 130 59	 84 57
ISHLT 0R **ISHLT 1R** **ISHLT 2/3R**	312 51 30	271 77 8	130 51 8	57 24 60

Table 1: Clinical characteristics of all cohorts. N/A Not available

Figures

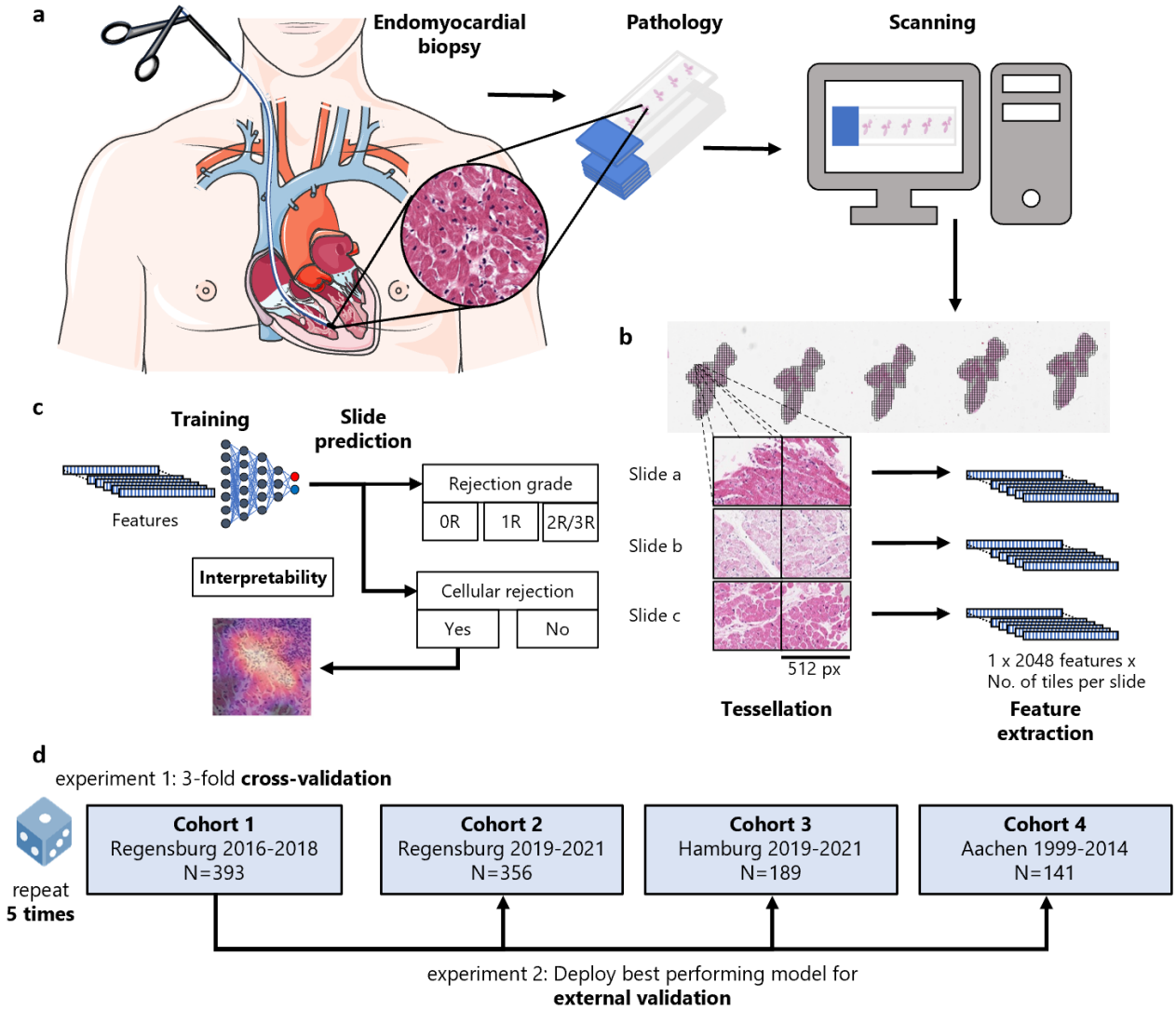


Figure 1: Outline of the study procedures. a) Routine endomyocardial biopsies of the right interventricular septum were taken from heart transplanted patients. These biopsies were then prepared into H&E stained histopathology slides, before being digitized and turned into whole slide images (WSIs) by use of a slide scanner. (Icon from smart.servier.com) **b)** To make these WSIs processable for our attention-based deep learning models, in a first step they need to be cut into smaller tiles while the background and artifacts are removed (tessellation). In the next step feature maps are extracted from all tiles from all slides using a publicly available neural network, which has been pre-trained by self-supervised learning with thousands of histopathology images. **c)** The resulting bags of feature maps per slide, together with expert pathologists' opinion on the occurrence of rejection on a slide level as target label, are then used as training input for an attention-based deep learning model. **d)** In a first experiment, three-fold cross-validation is performed within Cohort 1 and repeated 5 times. In a second experiment the best performing model from experiment 1 is externally validated on Cohorts 2, 3 and 4.

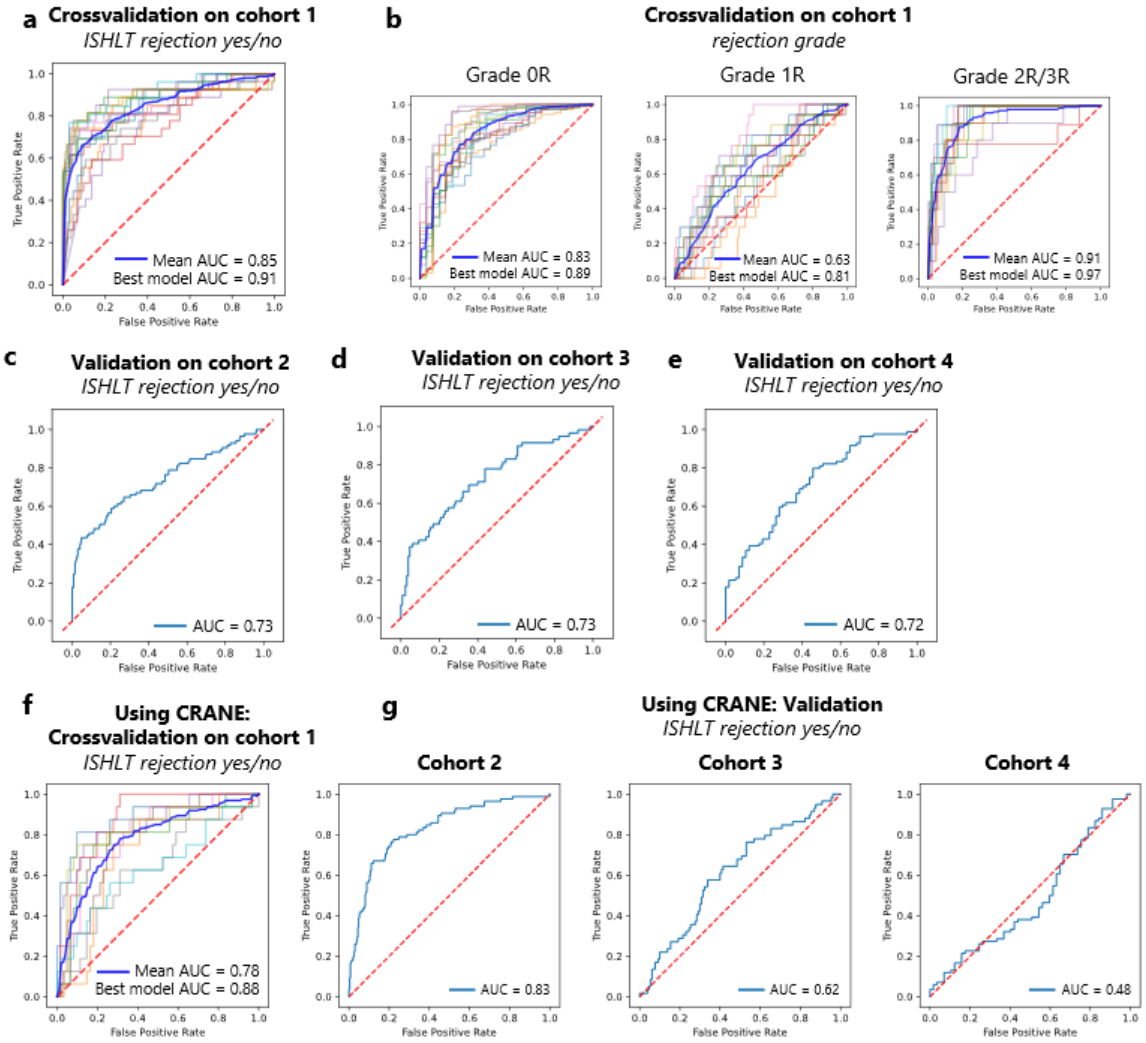


Figure 2: Deep learning can predict rejection and rejection grade from pathology images. Receiver operator characteristic curves (ROC) and mean area under the receiver operator curve. as measure of performance of the classifier for heart transplant rejection following 2004 revision of the International Society for Heart and Lung Transplantation (ISHLT) grading system. Showing binarized prediction (ISHLT rejection yes/no) (**a, c, d,** and **e**) and rejection grade (ISHLT 0R, 1R, 2/3R) (**b**) for cross-validation (**a** and **b**) and external validation (**c**, **d**, and **e**) experiments, as well as cross validation (**i**) and external validation (**j**) for binarized prediction (ISHLT rejection yes/no) using CRANE algorithm.

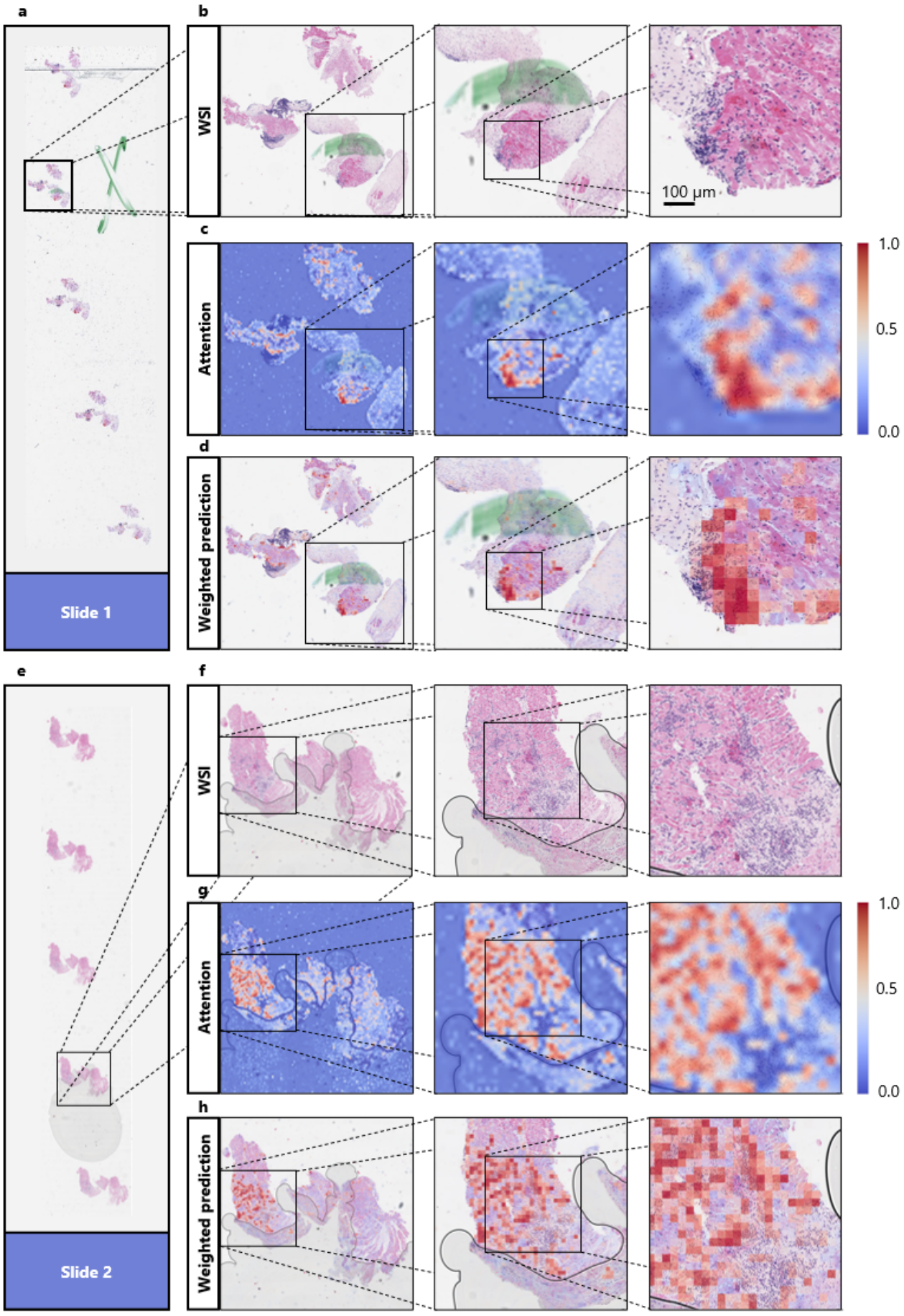
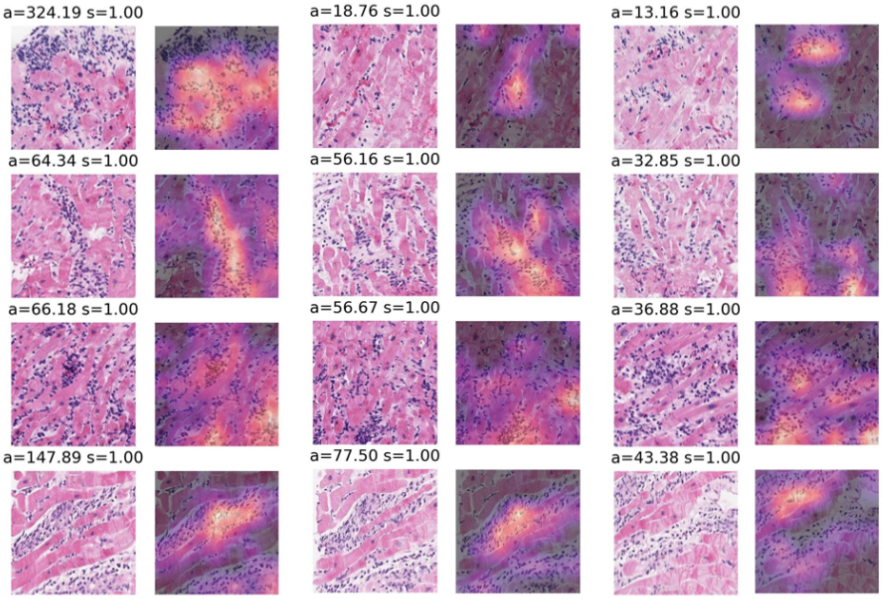


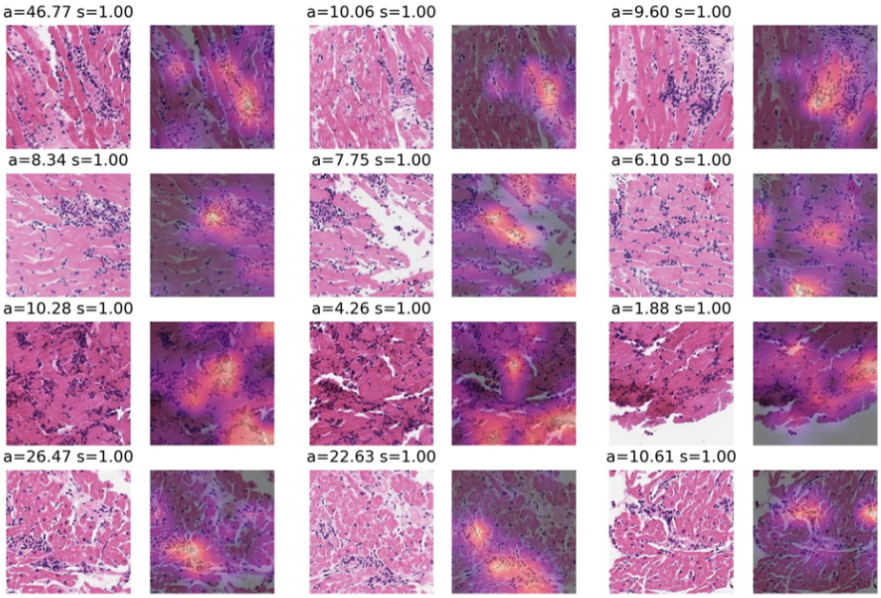
Figure 3: Explaining the models' decisions by visualizing the model's high attention regions. Different zoom levels of areas of the whole-slide-images (**b** and **f**) containing one patch of the

endomyocardial biopsy of two different slides (**a** and **e**) together with the attention-based heatmap of the corresponding slide region (**c** and **g**) and a heatmap showing the attention scores multiplied by the prediction scores (**d** and **h**). In attention-based heatmaps dark red indicates regions with a high attention, while dark blue indicates regions with a low attention [see scale in **c)** and **g**]. The network is focussing on areas of the whole slide image containing tissue, ignoring artefacts, like air bubbles and pen marks (**d** and **h**). The network was trained on cohort 1 for the binarized target (rejection yes/no) and deployed on cohort 2. For those two slides, the network was the “the most confident” about its decision (reflected by highest attention and prediction scores). The network is highlighting regions with a high number of lymphocytes between heart muscle tissue.

a Validation on cohort 2: *ISHLT rejection yes*



b Validation on cohort 3: *ISHLT rejection yes*



c Validation on cohort 4: *ISHLT rejection yes*

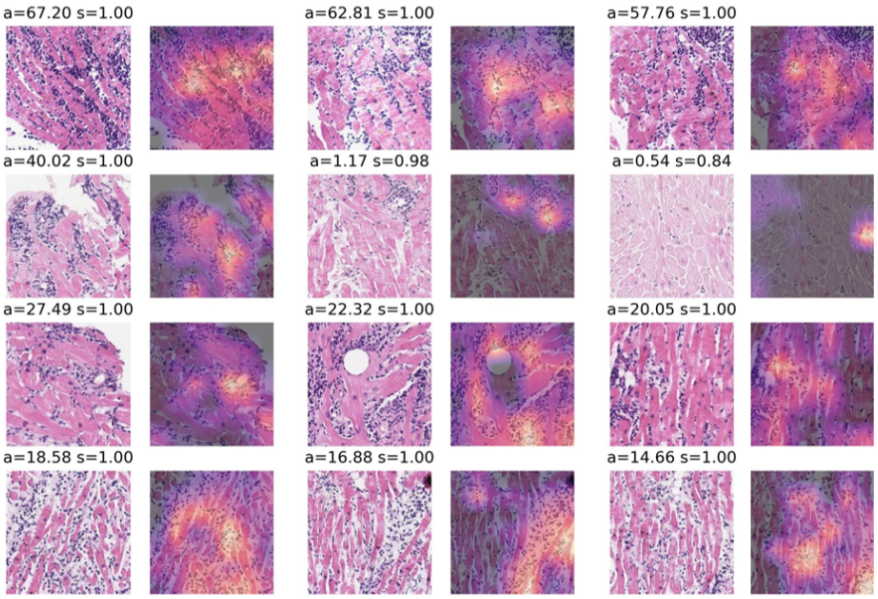


Figure 4: Explaining the models' decisions by visualizing the model's high attention 512x512 tiles. The three tiles (columns) with the highest average attention and prediction scores (attention = a, prediction score = s) for the four slides(rows) with the highest average prediction scores when deploying the best performing model to detect rejection (rejection yes/no) on the three test cohorts (**a**,**b**, and **c**). Together with the corresponding Grad-CAM images showing the network's spatial attention for each of the tiles. Regions with higher attention are yellow, while regions with low attention are in dark purple. The top tiles contain many lymphocytes infiltrating the myocytes, while the network's attention also appears to be lying on these immune cells.