Mapping and Diagnosing Augmented Whole Slide Image Datasets with Training Dynamics

Wenqi Shi, Benoit Marteau, May D. Wang
Georgia Institute of Technology
{wshi83, bmarteau3, maywang}@gatech.edu

Abstract

Pediatric heart transplantation represents the standard of care for children confronting end-stage heart failure. One of the most common postoperative complications, heart transplant rejection, has been monitored via surveillance endomyocardial biopsies and manual assessment by cardiac pathology experts. However, manual annotations with interobserver and intraobserver variability among cardiovascular pathology experts lead to significant disagreements about the severity of rejection. Artificial intelligence (AI)-enabled computational pathology usually requires large-scale manual annotations of gigapixel whole-slide images (WSIs) for effective model training. To address these challenges, we develop an AI-enabled rare disease detection framework for automating heart transplant rejection detection from WSIs of pediatric patients. Specifically, we conduct dataset cartography with data maps and training dynamics to map and diagnose the augmented samples, exploring the model behavior on individual instances during model training. Extensive experiments on internal and external patient cohorts have demonstrated the feasibility of both tile-level and biopsy-level detection with augmented samples. The proposed data-efficient learning framework may support seamless scalability to real-world rare disease detection without the burden of iterative expert annotations.

1 Introduction

EndoMyocardial Biopsy (EMB) screening is the standard care in detecting heart rejections following transplantation [10]. However, the manual interpretation of EMBs is subject to interobserver and intraobserver variability [14], leading to inconsistency in diagnosis and prognosis. Artificial Intelligence (AI)-enabled clinical decision support systems have advanced the objective and automated assessment of EMBs for improving procedure reproducibility and patient operative outcomes [10, 8, 9, 3, 18]. Recent investigations [8, 4] have highlighted the potential of AI models to assist human experts across a wide range of diagnostic tasks, including heart rejection detection. However, prior endeavors [4, 11] have encountered a primary challenge of small datasets with limited manual annotation of gigapixel Whole-Slide Images (WSIs), leading to poor domain adaptation. In addition, with major disagreements about the severity of rejection, learning from noisy labels is a challenge in computer-aided image analysis for complex WSI applications. In this study, we propose a rare disease detection framework to automate heart transplant rejection detection from WSIs for pediatric patients (Figure 1). We implement data augmentation via different generative models to facilitate data-efficient learning. Specifically, we leverage training dynamics via data map to map and diagnose both original and synthetically augmented training instances, exploring the behavior of the tile-level classification model on individual instances during the training process.

1st Workshop on Deep Generative Models for Health at NeurIPS 2023.

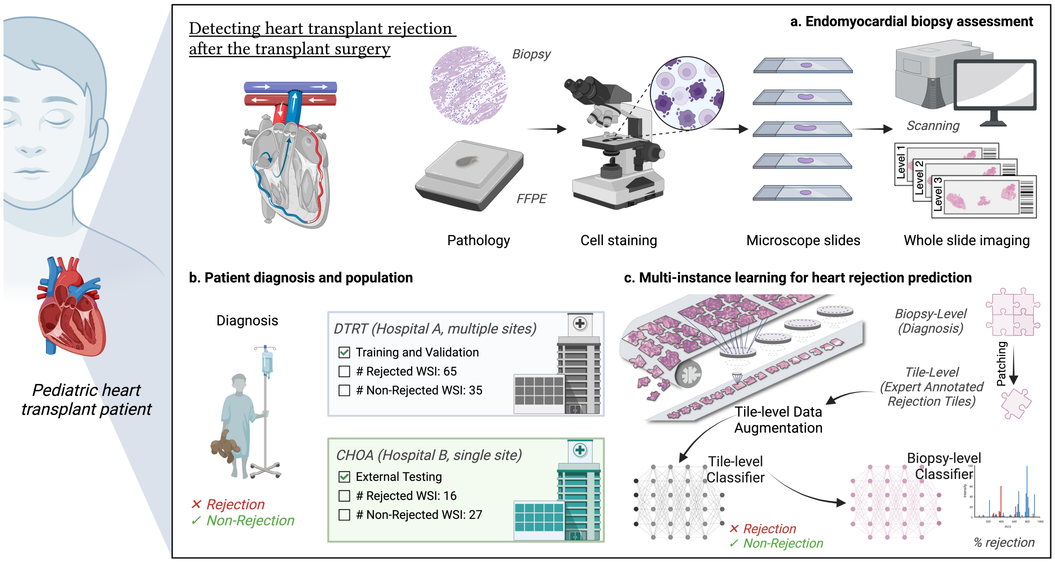

Detecting heart transplant rejection after the transplant surgery
a. Endomyocardial biopsy assessment
Biopsy
FFPE
Pathology
Cell staining
Microscope slides
Scanning
Level 1
Level 2
Level 3
Whole slide imaging
Pediatric heart transplant patient
b. Patient diagnosis and population
Diagnosis
DTRT (Hospital A, multiple sites)
☑ Training and Validation
☐ # Rejected WSI: 65
☐ # Non-Rejected WSI: 35
CHOA (Hospital B, single site)
☑ External Testing
☐ # Rejected WSI: 16
☐ # Non-Rejected WSI: 27
✗ Rejection
✓ Non-Rejection
c. Multi-instance learning for heart rejection prediction
Biopsy-Level (Diagnosis)
Patching
Tile-Level (Expert Annotated Rejection Tiles)
Tile-level Data Augmentation
Tile-level Classifier
✗ Rejection
✓ Non-Rejection
Biopsy-level Classifier
% rejection

Figure 1: Overview of the proposed framework for pediatric heart transplant rejection detection. Our proposed method first segments tissue regions in the WSI, patching them into smaller tiles. Considering the rare condition of heart transplant rejection, we employ advanced image generation approaches to augment rejection tiles for tile-level classification. Subsequently, the probabilities of tile-level rejection are used to train a biopsy-level rejection classifier.

2 Related Works

Advances in digital pathology have enabled potential adoptions of AI-enabled clinical decision support systems in analyzing gigapixel WSIs for objective diagnosis, prognosis, and therapeutic-response prediction [18, 8, 9, 4, 10, 11]. Seraphin et al. [18] demonstrated the feasibility and effectiveness of leveraging attention-based deep learning in predicting the degree of cellular rejection from pathology slides, with the degree of rejection defined by the International Society for Heart and Lung Transplantation grading system. With a total of 1,079 histopathology slides collected from 325 patients in three German transplant centers, researchers achieved an Area Under the Receiver Operating Characteristic (AUROC) range of 0.716 to 0.734 within external validation cohorts. In addition, they employed an explainable AI approach, GradCAM [17], to identify the spatial distribution of the attention layer for identifying potential pitfalls. The model interpretation outcomes suggested that deep learning models might primarily focus on lymphocytes and could potentially be misled by the occurrence of a quilty lesion. Similarly, Lipkova [8] presented a deep learning-enabled automated assessment of gigapixel WSIs obtained from EMBs for rejection and Quilty B lesions detection. They achieved accurate and robust assessment results with an AUROC ranging from 0.839 to 0.852 on external validation sets. In histopathology studies for other diseases, Lu et al. [9] advanced a deep-learning framework based on weak supervision to facilitate the automatic detection of subregions with high diagnostic values for WSI-level classification. Applications on the subtyping of renal cell carcinoma and non-small-cell lung cancer, as well as the detection of lymph node metastasis, demonstrated the effectiveness of data-efficient learning in the localization of morphological features on WSIs.

Given the rare conditions of heart transplant rejection, particularly in pediatric patients, existing studies usually suffer from limited sample sizes for model training and optimization. For example, Giuste et al. [4] endeavored to develop a deep learning model to automate the quantification of rejection risk using digitized images of biopsied tissues. Due to the small number of samples with expert annotations, researchers augmented training samples of around 1,100 tiles from 24 pediatric patients with GANs-generated images. Similarly, Mirzazadeh et al. [11] conducted inspirational image generation with rejection reference images to expand the pool of training samples, thereby facilitating the development of data-efficient models. Existing studies [4, 11] have demonstrated the effectiveness of leveraging generative models to augment limited training samples and improve prediction accuracy in pediatric heart transplant rejection detection. However, these findings remain preliminary based on a retrospective assessment in a small pediatric patient cohort with potential pitfalls. Firstly, given synthetic images from the same distribution as the original, existing models [4, 11] may suffer from generalization, especially without external validations. Secondly, there is a lack of effective evaluation of augmented samples. Considering existing evaluation metrics of generated images focusing more on efficiency, similarity, or diversity [20], it is insufficient to examine the role of generated samples in boosting data-efficient learning model performance. Thirdly, given the rare conditions and complicated patterns, annotation quality issues exist in the pixel-level manual annotation of gigapixel WSIs [10].

3 Methodology

Given the rarity of heart transplant rejection, we employ two different generative pipelines: (1) a combination of Progressive GAN (PGAN) [7] and Inspirational GAN (IGAN) [16], and (2) diffusion model [6], to augment the rejection samples in the training set. These augmented training samples fine-tune a pre-trained tile-level classifier to distinguish between rejection and non-rejection cases. Specifically, we leverage data cartography with training dynamics to minitor the optimization process with augmented images. We then train a separate biopsy-level classifier to estimate the rejection grade with tile-level probabilities as input to detect heart transplant rejection at the biopsy level.

PGAN and IGAN We combine two variations of GANs for high-quality tile generation. Initially, we leverage a PGAN [7] to first generate low-resolution images and progressively increase the size of the output image. Following PGAN training on all rejection and non-rejection training tiles via unconditional training, we then implement the IGAN [16] to generate synthetic rejection-specific tiles. IGAN enables the creation of a synthetic image closely aligned with a chosen image by identifying optimal parameters within the latent space. This is achieved by computing the distance between the features of the selected image and those of the GAN output using a pre-trained VGG-19 model as a feature extractor. Formally, a typical GAN usually contains a Generator *G* and a Discriminator *D*, along with a random sample vector by z. For the optimization of generative models, we employ the gradient-free Discrete One Plus One (DOPO) [1] optimizer for high-quality image generation. We then optimize the generation process by the Wasserstein GAN with Gradient Penalty (WGAN-GP) loss [5]. Given an input real image *x*, the WGAN-GP objective function is defined as:

$$\mathcal{L} = \min_{G} \max_{D} \mathbf{E}[D(x)] - \mathbf{E}[D(G(z))] + \lambda \mathbf{E}[(||\nabla_D(\tilde{x})||_2 - 1)^2], \tag{1}$$

where $\tilde{x} = x \cdot t + (1-t) \cdot G(z)$ denotes a combination of both a real and synthetic image, with $t \in [0, 1]$ representing the proportion of each in the combination. In the scenario of GANs, the Discriminator D often proves easier to train than the Generator G, leading to an imbalance that can result in vanishing gradients. The Gradient Penalty term is included in the objective function to alleviate this vanishing and exploding gradient problem.

Diffusion Models We generate synthetic rejection tiles utilizing the state-of-the-art diffusion models. Inspired by non-equilibrium thermodynamics, diffusion models establish a Markov chain of diffusion steps to progressively introduce random noise into the data [6]. The model then learns to reverse this diffusion process, thereby generating desired data samples from the noise:

$$q(x_1, ..., x_T | X_0) = \prod_{t=1}^{T} q(x_t | x_{t-1}) \tag{2}$$

with

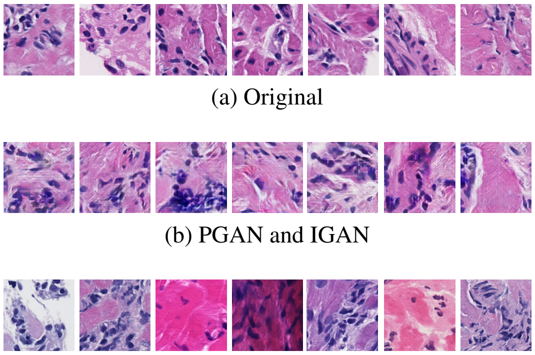
$$q(x_t | x_{t-1}) = \mathrm{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t \mathbf{I}) \tag{3}$$

where q denotes the data distribution, x_t denotes the latent space, and $\beta_t \in (0, 1)$ represents the variance of the noising process. With the exact reverse distribution $q(x_{t-1}|x_t)$, we can then generate synthetic image by sampling $x_T \sim \mathrm{N}(0, \mathbf{I})$ and perform the reverse denoising process. Specifically, we employ a conditional image generation approach using guided diffusion. This process incorporates a classifier to enable conditional training, focusing solely on generating rejection tiles.

Data Cartography Formally, consider a training dataset of size $N, \mathcal{D} = \{(\boldsymbol{x}, y^*)_i\}_{i=1}^{N}$, where the ith instance consists of the observation $\boldsymbol{x}_i$ and its corresponding ground truth label y_i^*. When minimizing empirical risk, we assume the model defines a probability distribution over labels given an observation. For a stochastic gradient-based optimization, the model involves random ordering of the training instances across E epochs during each epoch. We then define the training dynamics of instance i across the E epochs. Initially, we capture the confidence of the model in assigning the true label to an observation $\boldsymbol{x}_i$ based on its probability distribution. We define the model confidence $\hat{\mu}_i$ as the average probability of the true label y_i^* across E epochs: $\hat{\mu}_i = \frac{1}{E}\sum_{e=1}^{E} p_{\boldsymbol{\theta}^{(e)}}\left(y_i^* \mid \boldsymbol{x}_i\right)$, where $p_{\theta^{(e)}}$ denotes the probability of the model and $\theta^{(e)}$ represents the model parameters at the end of the eth epoch. Secondly, as a more intuitive statistic, we define correctness $\hat{\phi}_i$ as the fraction of times the model correctly labels instance x_i across epochs E: $\hat{\phi}_i = \frac{1}{E}\sum_{i=1}^{E} \mathbb{1}\left(\hat{y}_i = y_i^* \mid x_i\right)$. Note that correctness can only have $1 + E$ discrete values. Lastly, we further consider variability $\hat{\sigma}_i$ quantifying the spread of $p_{\boldsymbol{\theta}^{(e)}}\left(y_i^* \mid \boldsymbol{x}_i\right)$ across E epochs using the standard deviation of confidence: $\hat{\sigma}_i = \sqrt{\frac{\sum_{e=1}^{E}\left(p_{\boldsymbol{\theta}^{(e)}}\left(y_i^* \mid \boldsymbol{x}_i\right) - \hat{\mu}_i\right)^2}{E}}$. We then leverage confidence, variation, and correctness to visualize training dynamics using data map [19]. A given training sample that receives a consistent label assignment from the model, irrespective of accuracy, represents low variability. Conversely, if the model struggles with labeling (i.e., prediction) throughout the training process, this instance represents high variability. Training dynamics can be evaluated at varying steps and epochs with metrics obtained above as the coordinates in the data map.

4 Experiment and Discussion

In the experiments, we first performed quantitative and qualitative analysis to evaluate generation quality. Compared to the original images, we observed very similar patterns in synthetic images from both GANs and diffusion. Specifically, synthetic images from diffusion models cover more diverse patterns. For qualitative evaluation, we presented examples of synthetic images generated by GANs and diffusion models in Figure 2. The quantative results indicate that diffusion could generate more diverse synthetic tiles while remaining closer to the distribution of real images, with a higher Inception Score of 3.67, a lower sFID score of 97.6, a higher precision score of 0.61, and a higher recall of 0.69.


(a) Original

(b) PGAN and IGAN

(c) Diffusion

Figure 2: Examples of (a) original tiles and generated tiles using (b) GANs and (c) diffusion model.

Extensive experiments on internal and external patient cohorts demonstrate the feasibility of both tile-level and biopsy-level detection with augmented samples (Appendix B), with the best AUROC of 0.9984 and 0.7681, respectively. Experimental results suggest that augmenting training samples could boost model performance by increasing sample size or solving dataset imbalance in biopsy-level classifications. For tile-level classification, our experiments reveal a distinct pattern under different experimental settings. Although data augmentation enhances model performance on the internal set, as evidenced in previous studies [4, 11], its effect on the external set is less promising. This might be attributed to generative model collapse, distribution shift, or overfitting issues. For biopsy-level model interpretation, Figure 3 represents the rejection probabilities, with red indicating a higher risk for rejection signs. The interpretation results for tile-level models can be found in Appendix C.

In the generation of data maps, we incorporated all epochs into the calculation of training dynamics, beginning with the initial epoch. Figure 4b and 4a provide examples of data maps for the augmented training set based on a ResNet152 classifier by diffusion and GANs, respectively. According to the measurements, the top-left corner of the data map, characterized by low variability and high confidence, populates the *easy-to-learn* examples, which form the majority of the original dataset.

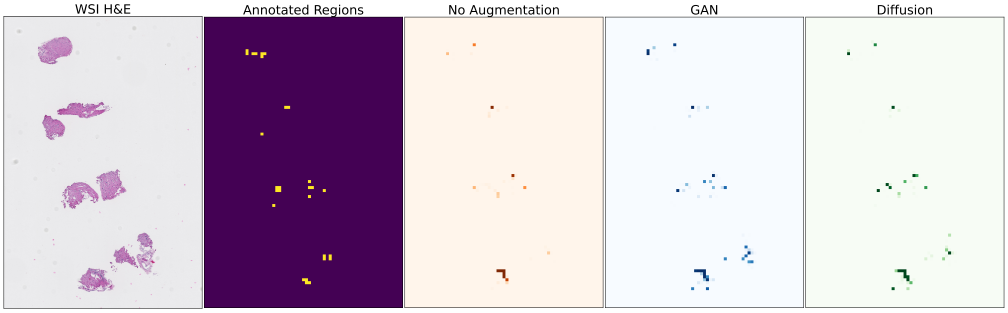
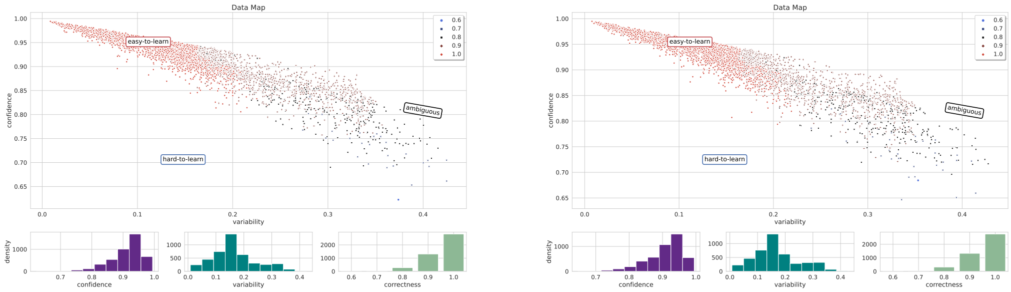


Figure 3: Comparisons among annotated important regions containing signs of cellular rejection and predicted heatmaps with and without augmentations.



(a) GANs-Augmented

(b) Diffusion-Augmented

Figure 4: Data map for augmented training set based on a ResNet152 classifier. The x-axis shows variability, the y-axis shows confidence, and the colors indicate correctness. The top-left corner of the data map (low variability, high confidence) corresponds to easy-to-learn examples, the bottom-left corner (low variability, low confidence) corresponds to hard-to-learn examples, and examples on the right (with high variability) are ambiguous. Density plots for the three measures based on training dynamics are shown towards the bottom.

The examples with high variability, located on the right side of the map, are inherently *ambiguous* and represent complex patterns present in some tile-level instances. Conversely, the *hard-to-learn* examples can be found in the bottom-left corner, defined by low variability and low confidence. As suggested by Figure 4, these *hard-to-learn* samples are scarce, indicating the relatively high quality of generated samples. Interestingly, more "*ambiguous*" data points are observed in the diffusion-based map compared to the one using GANs. This suggests a more diverse sample distribution that could potentially enhance model performance. Training dynamics of different augmentations could potentially be utilized to evaluate and select augmented data samples. Furthermore, we investigated the feasibility of creating data maps at various epochs to understand the evolution of training dynamics over time (Appendix D). Data maps facilitate us to intuitively identify the contribution of individual samples when monitoring the optimization process and evaluating augmented samples.

5 Conclusion

In this study, we developed and validated AI-enabled clinical decision support to automate heart transplant rejection detection for pediatric patients based on data augmentation of limited patient samples. With dataset cartography using data maps and training dynamics, we qualitatively mapped and diagnosed the contribution of augmented samples, exploring the behavior of the model on individual instances during model optimization for limited samples in rare diseases. Extensive experiments on internal and external patient cohorts demonstrated the feasibility and effectiveness of both tile-level and biopsy-level detection. Explainable AI results improved model interpretability via high-resolution heatmaps, enabling human intervention and clinical validation. The proposed data-efficient learning framework mau facilitate seamless scalability in detecting rare diseases without imposing the burden of iterative expert annotations, thereby increasing effectiveness and efficiency.

References

[1] J. Cai and G. Thierauf. Evolution strategies for solving discrete optimization problems. *Advances in Engineering Software*, 25(2):177–183, 1996. Computing in Civil and Structural Engineering.

[2] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[3] M. Colvin-Adams and A. Agnihotri. Cardiac allograft vasculopathy: current knowledge and future direction. *Clinical transplantation*, 25(2):175–184, 2011.

[4] F. O. Giuste, R. Sequeira, V. Keerthipati, P. Lais, A. Mirzazadeh, A. Mohseni, Y. Zhu, W. Shi, B. Marteau, Y. Zhong, et al. Explainable synthetic image generation to improve risk assessment of rare pediatric heart transplant rejection. *Journal of Biomedical Informatics*, 139:104303, 2023.

[5] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans, 2017.

[6] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[7] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[8] J. Lipkova, T. Y. Chen, M. Y. Lu, R. J. Chen, M. Shady, M. Williams, J. Wang, Z. Noor, R. N. Mitchell, M. Turan, et al. Deep learning-enabled assessment of cardiac allograft rejection from endomyocardial biopsies. *Nature medicine*, 28(3):575–582, 2022.

[9] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

[10] F. Mahmood and E. J. Topol. Digitising heart transplant rejection. *The Lancet*, 400(10345):17, 2022.

[11] A. Mirzazadeh, A. Mohseni, S. Ibrahim, F. O. Giuste, Y. Zhu, B. M. Shehata, S. R. Deshpande, and M. D. Wang. Improving heart transplant rejection classification training using progressive generative adversarial networks. In *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4. IEEE, 2021.

[12] M. B. Muhammad and M. Yeasin. Eigen-cam: Class activation map using principal components. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[13] H. G. Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 983–991, 2020.

[14] M. Reid-Nicholson, R. Kulkarni, B. Adeagbo, S. Looney, and J. Crosby. Interobserver and intraobserver variability in the calculation of the lipid-laden macrophage index: implications for its use in the evaluation of aspiration in children. *Diagnostic Cytopathology*, 38(12):861–865, 2010.

[15] M. E. Richmond, S. D. Zangwill, S. J. Kindel, S. R. Deshpande, J. N. Schroder, D. P. Bichell, K. R. Knecht, W. T. Mahle, M. A. Wigger, N. A. Gaglianello, E. Pahl, P. M. Simpson, M. Dasgupta, P. E. North, M. Hidestrand, A. Tomita-Mitchell, and M. E. Mitchell. Donor fraction cell-free dna and rejection in adult and pediatric heart transplantation. *The Journal of Heart and Lung Transplantation*, 39(5):454–463, 2020.

[16] B. Rozière, M. Riviere, O. Teytaud, J. Rapin, Y. LeCun, and C. Couprie. Inspirational adversarial image generation. *IEEE Transactions on Image Processing*, 30:4036–4045, 2021.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[18] T. P. Seraphin, M. Luedde, C. Roderburg, M. van Treeck, P. Scheider, R. D. Buelow, P. Boor, S. H. Loosen, Z. Provaznik, D. Mendelsohn, et al. Prediction of heart transplant rejection from routine pathology slides with self-supervised deep learning. *European Heart Journal-Digital Health*, 4(3):265–274, 2023.

[19] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, and Y. Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *arXiv preprint arXiv:2009.10795*, 2020.

[20] A. Vahdat, K. Kreis, and J. Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021.

[21] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.

A Data Description

In this study, pediatric heart biopsies were gathered from two institutions: (1) the multi-center prospective blinded study, DNA-Based Transplant Rejection Test (DTRT) [15], and (2) Children's Healthcare of Atlanta (CHOA). All experiments were conducted in compliance with relevant guidelines and regulations, with informed consent obtained from all participants. Biopsy-level annotations were acquired directly from clinicians at the source institutions. For tile-level annotations, the ground truth of Acute Cellular Rejection (ACR) regions in WSIs was obtained from clinical experts using HistomicsTK[1]. Following multi-instance learning formulation, training tiles were labeled as 'rejection' if they overlapped more than 60% with an annotated region. An example of biopsy- and tile-level images can be seen in Figure 5. For both tile- and biopsy-level model development, a train and test split was performed at the patient level, with the detailed distributions documented in Table 1. The training and validation sets were drawn from DTRT, while the external testing set was obtained from CHOA. This setup ensures a robust evaluation across varied datasets.

For data pre-processing and quality control, the raw materials were WSIs sourced from DTRT and CHOA. Given the large and variable sizes of WSIs, we performed a two-stage pre-processing before feeding them into a computer-aided image analysis system. Initially, we segmented the WSIs to separate the tissue from the background. This segmentation was achieved using Otsu's thresholding method on a downscaled version of the WSI, where the resolution was reduced by a factor of 10. Following this, we generated non-overlapping tiles from the segmented WSI, each measuring 256×256 pixels at 40X magnification. This 40X magnification was chosen because it allowed multiple muscle cells and white blood cells to be included within each tile, which is crucial for detecting signs of ACR. We retained only tiles that overlapped with more than 80% of the previously identified tissue for quality control.

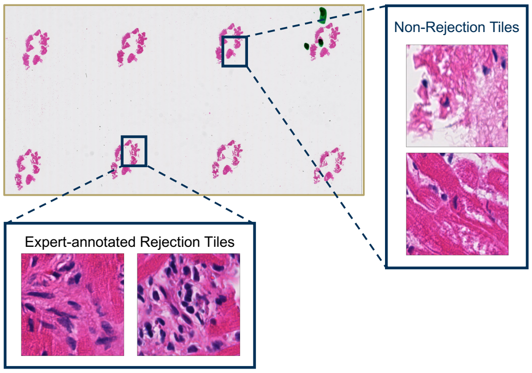


Figure 5: Illustration of biopsy- and tile-level images annotated by clinical experts for the identification of regions exhibiting cellular rejection indicators. Local regions characterized by signs of cellular rejection are extracted for automating rejection detection.

[1] HistomicsTK: `https://github.com/DigitalSlideArchive/HistomicsTK`

Table 1: Summary of tile-level and biopsy-level patient data: training and validation sets derived from DTRT, and external testing set sourced from CHOA.

Tile-Level	**Training (DTRT)**	**Testing (CHOA)**
Rejection	2330	105
Non-Rejection	3256	12167
Biopsy-Level	**Training (DTRT)**	**Testing (CHOA)**
Rejection	65	16
Non-Rejection	35	27

B Tile- and Biopsy-Level Classification Results

B.1 Tile-level Classification

Tables 2 reports the AUROC scores on both external and internal testing sets for heart transplant rejection classification. The performance of four different convolutional neural network models (VGG19, ResNet50, ResNet152, and DenseNet161) was evaluated, each using two distinct types of data augmentations: GAN-based and Diffusion-based. In the case of internal validation, the GAN-based data augmentation method achieved the highest AUROC scores across the VGG-19 (0.9884), ResNet152 (0.9806), and DenseNet161 (0.9984) models. This demonstrates that the GAN augmentation technique provided superior performance for these models when applied to the internal testing dataset. Conversely, when considering the external testing set (CHOA), the baseline models without any data augmentation surprisingly outperformed both the GANs and diffusion-based augmentations for the VGG-19, ResNet50, and ResNet152 models, with competitive AUROC scores of 0.9508, 0.9854, and 0.9863, respectively.

Table 2: Tile-level classification results (AUROC) on the external testing set (CHOA) and internal testing set (DTRT) for heart rejection classification.

Augment	**VGG19**	**ResNet50**	**ResNet152**	**DenseNet161**
External Testing Set (CHOA)				
Baseline	**0.9508**	**0.9854**	**0.9863**	0.8661
GAN	0.6748	0.9640	0.9849	**0.9717**
Diffusion	0.6133	0.9094	0.9421	0.8437
Internal Testing Set (DTRT)				
Baseline	0.9384	0.9422	0.9622	0.9950
GAN	**0.9884**	0.8963	**0.9806**	**0.9984**
Diffusion	0.9760	**0.9479**	0.9436	0.9895

B.2 Biopsy-level Classification

Tables 3 presents the biopsy-level classification results for heart transplant rejection on external and internal testing sets, respectively. For the internal testing set, the diffusion augmentation method consistently outperformed the others across all evaluated models, registering AUROC scores of 0.7297 for VGG-19, 0.7615 for ResNet50, 0.7451 for ResNet152, and 0.7681 for DenseNet161. On the external testing set (CHOA), the diffusion augmentation method once again delivered the highest AUROC scores for both the VGG-19 and ResNet50 models. However, for the ResNet152 model, the GAN augmentation method emerged as the most effective, achieving an AUROC score of 0.6551. These results highlight a significant discrepancy between internal and external validations at the biopsy level, thereby highlighting the importance of external validation in assessing the generalizability of models.

Table 3: Biopsy-level classification results on the external testing set (CHOA) and internal testing set (DTRT) for heart transplant rejection classification.

Augment	**VGG19**	**ResNet50**	**ResNet152**	**DenseNet161**
External Testing Set (CHOA)				
Baseline	0.4954	0.5150	0.4410	**0.4988**
GAN	0.4514	0.3519	**0.6551**	0.4676
Diffusion	**0.6481**	**0.5428**	0.5405	0.4606
Internal Testing Set (DTRT)				
Baseline	0.7077	0.7571	0.7132	0.7187
GAN	0.7198	0.7297	0.7330	0.7143
Diffusion	**0.7297**	**0.7615**	**0.7451**	**0.7681**

C Model Interpretation

For tile-level model interpretation, Figure 6 presents visual representations of model interpretation (i.e., heatmaps) using multiple explainable AI methods, including GradCAM++ [2], EigenCAM [12], ScoreCAM [21], and AblationCAM [13]. Model interpretation outcomes highlight important regions (red) in heatmaps, such as white blood cell infiltration and interstitial edema, for potential evidence of rejection region prediction. Similarity among annotated and important regions in both augmented and baseline heatmaps demonstrates that our proposed model could make predictions based on consistent evidence from clinical experts.

D Training Dynamics Details

We investigated the feasibility of creating data maps at various epochs before model convergence to understand the evolution of training dynamics over time (see Figure 7). From data map over epochs, we can notice the trend of instances moving from right to top-left, perfectly modeling the model training process. At convergence, model could correctly predict the majority of training instances with consistently high confidence.

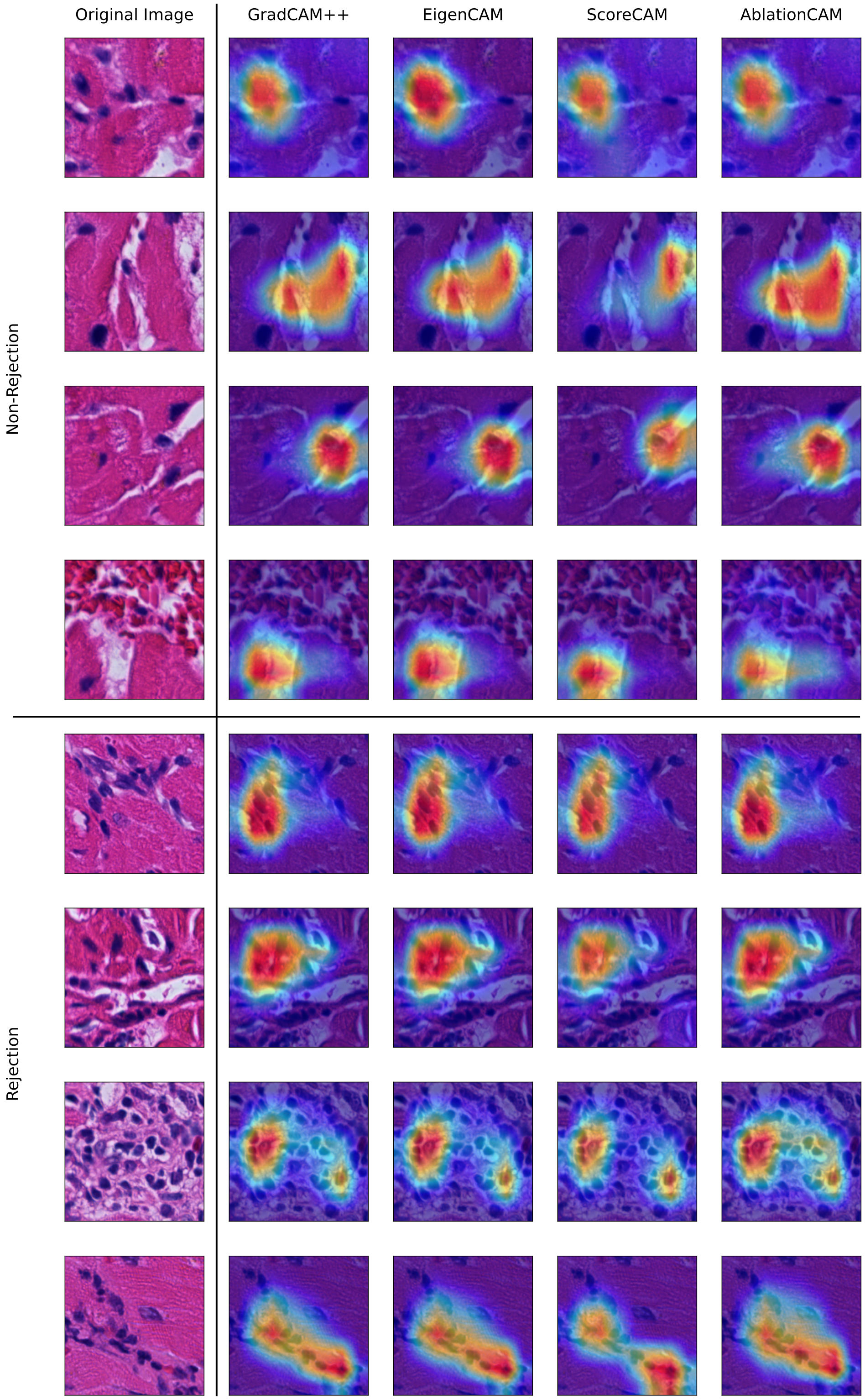
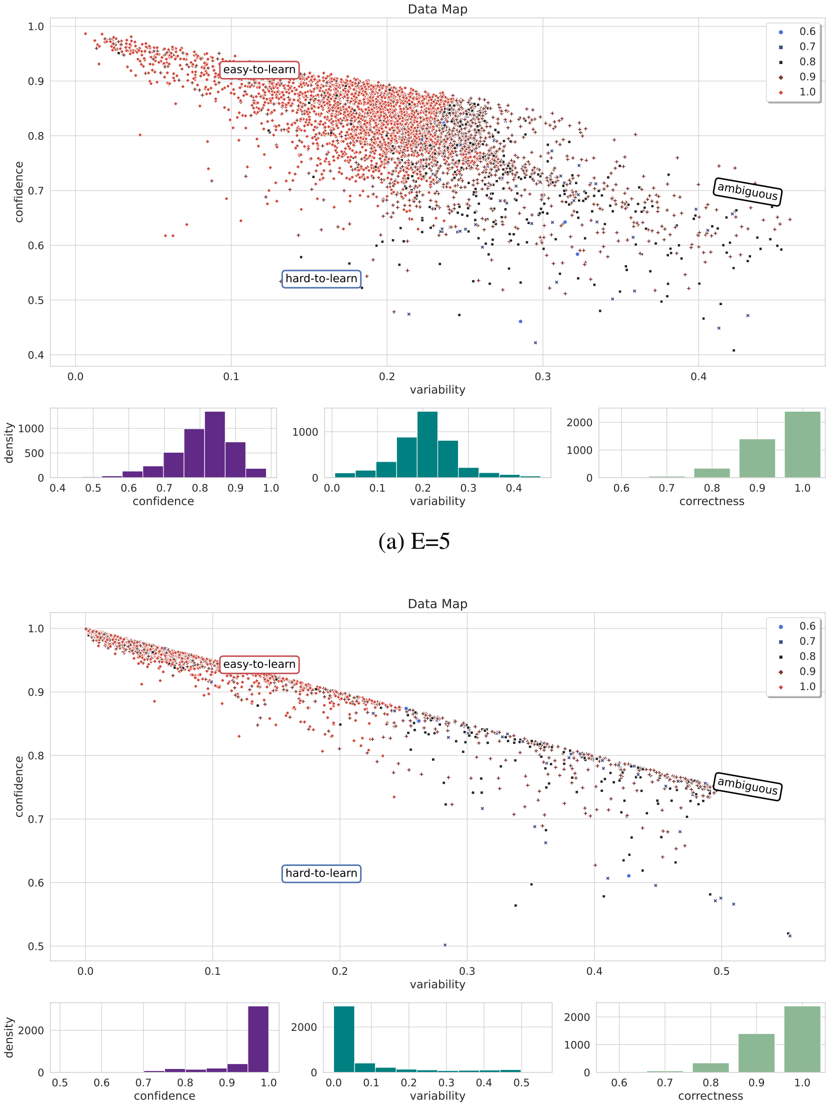


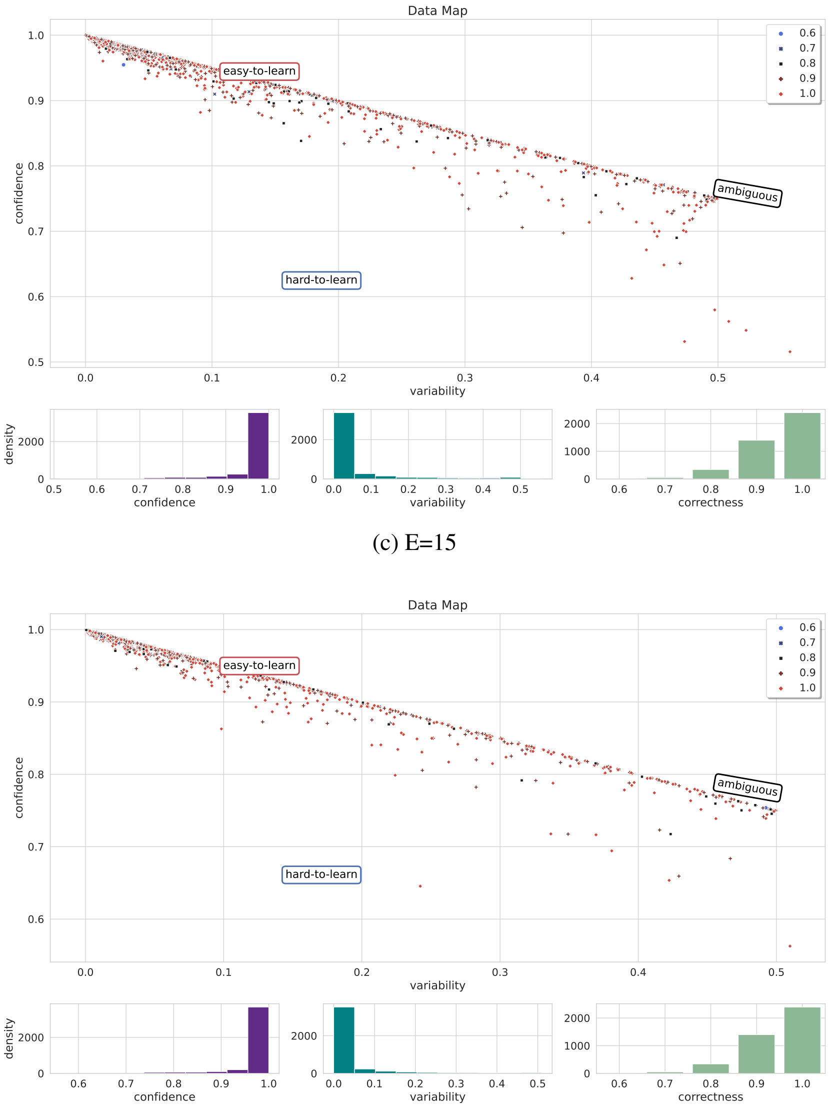
Figure 6: Model interpretation outcomes with important regions highlighted using Grad-Cam++, Eigen-CAM, Score-CAM, and Ablation-CAM for clinical validation and potentially novel biomarker identification.



(a) E=5

(b) E=10

Figure 7: Data map for original DTRT training set over 20 epochs with the same axes as Figure 4. Density plots for the three different measures based on training dynamics are shown towards the bottom.



(c) E=15

(d) E=20

Figure 7: Data map for original DTRT training set over 20 epochs with the same axes as Figure 4. Density plots for the three different measures based on training dynamics are shown towards the bottom. (Continued)