Predicting Heart Transplant Rejection from Longitudinal Cardiac Histology Images

Vivek Gopalakrishnan[1], Daniel Shao[1], Anurag Vaidya[1]

Abstract. Cardiac allograft rejection, a severe complication following heart transplant, occurs when a recipient's immune system attacks the transplanted heart. We explore statistical modeling (Hidden Markov Model) and neural network (Long Short Term Memory) methods for predicting a patient's risk of transplant rejection from longitudinal cardiac biopsies. We implement a Bayesian regression technique using HMMs which achieves a MSE of 0.07, beating simple linear regression baseline (MSE of 0.48). The addition of biopsy image data in our LSTM produces a model that is able to predict future rejection with an AUC of 0.96 ± 0.04.

1 Introduction

Approximately 3,000 heart transplants are performed annually in the United States alone, and an estimated 50-80% recipients experience at least one rejection episode [1]. Patients who receive heart transplants undergo routine cardiac biopsies to check for cellular indicators of rejection. However, under the current standard of care, diagnosing transplant rejection in a timely manner is difficult for two main reasons: (1) qualitative assessment of cardiac biopsies by pathologists is subjective and suffers from high inter-observer variability, and (2) pathologists typically predict rejection risk based solely on a patient's most recent biopsy, losing potentially valuable diagnostic information from previous biopsies [2]. In recent years, the field of *computational pathology* has endeavored to inform clinical decision making through the quantitative assessment of digitized histology images [3], often called whole slide images (WSIs). Computational pathology methods have proven successful in a variety of clinical tasks such as semantic segmentation of nuclei [4] and tissue classification [5].

In this work, we seek to develop computational pathology methods that forecast the risk of a heart transplant recipient rejecting their donor organ. Specifically, we ask, **if given a longitudinal set of cardiac histology images from a patient, can we predict their future risk of transplant rejection?** We first train a neural network to generate risk score for each WSI, then use a Hidden Markov model (HMM) to model the trajectory of risk scores. We also develop a Long Short Term Memory (LSTM) model to predict future rejection given longitudinal cardiac biopsies for a patient (Figure 3).

2 Methods

2.1 Data

Our dataset is comprised of post-transplant heart biopsies from 342 patients treated at the Brigham and Women's Hospital. Each patient underwent at least three cardiac biopsies at distinct time points after transplant, with a total 3,278 biopsies across all patients (Figure 1). These cardiac biopsies were stained with Hematoxylin and Eosin (the principal tissue stain used in histological medical diagnosis) and digitized to create WSIs. Each biopsy was labeled by a pathologist to indicate whether signs of immune rejection were present on the WSI. From these ground truth labels, we formally define our task as follows: For a patient who has undergone biopsies $X_1, \ldots, X_n$ with corresponding diagnoses $Y_1, \ldots, Y_n \in \{0, 1\}$, predict Y_n given $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_{n-1}, Y_{n-1})\}$ or some subset of $\mathcal{D}$. Through this setup, we simulate the clinical setting in which $n - 1$ biopsies have been performed, and we wish to identify the likelihood of future immune rejection at the time of the next biopsy.

2.2 Feature Extraction

Since the size of digitized WSIs is on the order of gigapixels, we used `CLAM` — a weakly supervised method for extracting features from regions of high diagnostic value — to extract clinically meaningful low-dimensional feature representations before proceeding with any supervised tasks. In `CLAM`, each WSI is deconstructed into non-overlapping 256 $\times$ 256 pixel sub-regions [6]. The third convolutional block of a ImageNet pretrained `ResNet-50` model is then used to extract a 1024-dimensional feature vector from each patch [7]. Consequently, each WSI is represented by a variable number of feature vectors.

[1] Department of Health Sciences and Technology, Harvard-MIT

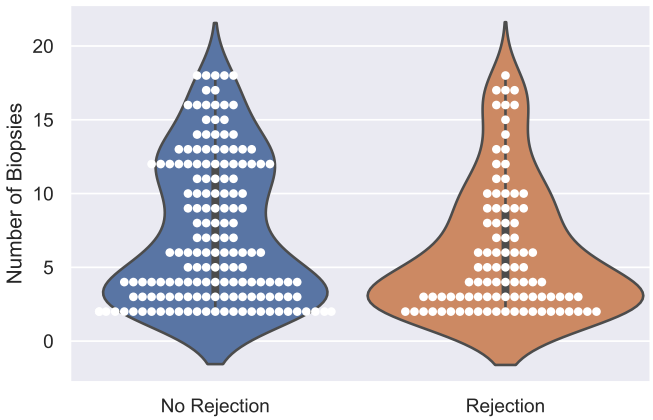

Figure 1: Patients with no sign of transplant rejection ($N = 203$) underwent a median 6 post-operative biopsies, compared to 4 biopsies for patients who did show signs of rejection ($N = 139$). Of the 342 patients examined, approximately 37% of patients showed signs of immune rejection in at least one of their biopsies

Our models require a single vector to serve as input for each time point. However, naively taking the unweighted mean of a WSI's feature vectors would lead to an uninformative embedding, since immune rejection is measured by the region with the *most* abnormal immune cell activity, rather than the average immune cell profile. To place higher weight on regions with abnormal immune cell activity, we use an attention-based neural network to learn attention scores such that feature vectors from regions of greater diagnostic value have higher weights. Implementation details for the network are discussed in Appendix C. We use these learned attention scores to compute a weighted average feature vector for each biopsy. Consider the WSI from the i'th biopsy of patient j. For m patches and corresponding feature vectors $x_1, \ldots, x_m \in \mathbb{R}^{1024}$ with attention scores $a_1, \ldots, a_m \in [0, 1]$ and $\sum_{i=1}^{m} a_i = 1$, our final feature vector is $\hat{x}_{i,j} = \sum_{k=1}^{m} a_k x_k$.

2.3 Statistical Modeling of Rejection Risk One natural approach for predicting rejection risk is to examine the trajectory of rejection likelihood from previous biopsies. We model the risk of rejection at each biopsy (i.e., each biopsy's *risk score*) as $R_{i,j} \in [0, 1]$, which represents the j-th patient's probability of rejection at the time of their i-th biopsy for $i = 1, \ldots, n_j$. To generate these risk scores, we trained an artificial neural network (ANN) which uses x_{ij} to predict the pathologist's diagnosis Y_{ij} at the time of biopsy i. The architecture of this ANN is covered in Appendix D. Because staining and imaging techniques can produce a high variance in WSI samples, we also implement adversarial attacks via the fast-gradient sign method while training to improve model robustness (details also in Appendix D). Let n_j be the total number of biopsies for patient j. Now, given a sequence of risk scores $R_{1,j}, \ldots, R_{n_j-1,j}$, our goal is to predict $R_{n_j,j}$. For this task, we elected to use a Hidden Markov Model (HMM), a popular algorithm in disease modeling due to its ability to capture underlying statistical patterns in disease progression[8].

HMMs are a statistical method for computing the probability of observed sequential data, making them ideal for modeling the trajectory of transplant rejection risk scores. In an HMM, we assume that there exists a set of discrete latent states $\{S_1, \ldots, S_K\}$ and that, at every time point, the subject is in exactly one of these states. In our analysis, this modeling choice translates to assuming that there exist unobserved clinical phenotypes that underlie the process of transplant rejection (e.g., if $K = 3$, then the states could represent *no rejection*, *rejection*, and a *transition* state where immune cells are actively being recruited to the heart, but have not yet started attacking the tissue). Additionally, in this model, the transition probabilities between states is governed by a first order Markov chain, with initial state

distribution $\pi_0 \in [0,1]^K$ and transition matrix $A \in [0,1]^{K\times K}$.[1] Finally, the observed quantity (i.e., the risk score) is sampled from an *emission distribution* specified by the latent state. In this analysis, we assume the emission distribution to be a univariate Gaussian, i.e., $R \mid S_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$.

We use a Bayesian modeling approach to estimate $R_{n_j,j}$. First, we split our data into a training and testing set. We use the training set (i.e., the full sequence of risk scores for a subset of patients) to estimate the parameters of the emission distribution using an iterative Expectation-Maximization (EM) algorithm with 1000 iterations. Next, given the emission distribution parameters and the testing data (only $R_{1,j}, \ldots, R_{n_j-1,j}$), we use the Viterbi algorithm to estimate the state distribution at the $n_j - 1$-th time point, $\hat{\pi}_{n_j-1}$. Finally, we treat $\hat{\pi}_n = A\hat{\pi}_{n_j-1}$ as the prior distribution over the emission distribution of $R_{n_j,j}$, and estimate $\hat{R}_{n_j,j}$ as the posterior mean of the conditional distribution $R_{n_j,j} \mid \hat{\pi}_n$.

Because each patient in our data set does not have the same number of biopsies, we use a linear interpolation scheme to temporally align the risk scores from all patients. Our interpolation methodology, which we also use when training our supervised models, is detailed in Appendix E.

2.4 Neural Network Prediction of Rejection Risk The HMM assumes that the current state depends only on the previous state. Standard recurrent neural networks are popular for modeling time series data, but also struggle to model long range dependencies in longitudinal data. To capture long range relationships, we use a modification of the standard RNN called the Long Short Term Memory (LSTM) model. Unlike standard RNNs, in which each new step depends only on the previous state and data from the new timepoint, LSTMs utilize gates to control how past data is propagated at each time step. Specifically, each step has a forget gate to control the information passed on from past states, an input gate to control how data from the current timepoint modifies the current state, and an output gate to control how the current state feeds into the next hidden state. The properties of these gates are learned by the model through parameters W_i, U_i, W_0, U_0, V_0. For the j'th layer of the LSTM at timepoint t, cell state c_t^j, the output h_t^j is given by $h_t^j = o_t^j tanh(c_j^t)$. The output gate is

$$o_t^j = \sigma(W_0 x_t + U_0 h_{t-1} + V_0 c_t)^j$$

The memory cell is updated as

$$c_t^j = f_i^j c_{t-1}^j + i_t^j \hat{c}_t^j$$

such that the past memory is partially forgotten and replaced by the new memory content based on input data x_t at time t.

$$\hat{c}_t^j = tanh(W_c x_t + U_c h_t - 1)^j$$

The forget gate f_t^j and input gate i_t^j are

$$f_t^j = \sigma(W_f x_t + U_f h_{t-1} + V_f c_{t-1})^j$$

$$i_t^j = \sigma(W_i x_t + U_i h_{t-1} + V_i c_{t-1})^j$$

The input to our LSTM is a matrix of size $n_j - 1 \times 1024$, where column i is the attention-weighted biopsy embedding $\hat{x}_{i,j}$. Our LSTM has a single hidden layer. A fully connected layer with 2 output nodes was then used to perform classification. The LSTM's parameters were initialized by a uniform random distribution. The model was trained for 600 epochs with cross entropy loss, a learning rate of 1 (exponentially decaying by a factor of 0.5 every 50 epochs), and the SGD optimizer (with momentum of 0.45).

For the model construction procedure $M(x)$ described above, where x is the input data to our model, we construct five baseline models to understand how the inclusion of time-series data our model's ability to predict Y_{n_j}. These models were $M(\hat{x}_1)$, $M(\hat{x}_{n_j-1})$, $M(\hat{x}_1, \hat{x}_{n_j-1})$, $M(\hat{x}_1, \ldots, \hat{x}_{n_j-1})$, and $M(\tilde{x}_1, \ldots, \tilde{x}_{n_j-1})$ where $\tilde{x}_i$ is the concatenation of $\hat{x}_i$ and y_i. Five fold stratified cross validation

[1] To make the Markov model a valid probabilistic process, the additional constraints on the parameters are $\|\pi_0\|_1 = 1$ and $\sum_{j=1}^K A_{ij} = 1$ for all $i \in [K]$.

was performed for each experiment (80% data for training, 20% data for testing). For each epoch, the model giving the best area under the curve (AUC) score for used for testing. Class-weighted averages of precision, recall, and F1 score were used to compare model performance. All modeling was done on three GeForce RTX 2080 Ti.

3 Results

3.1 Risk Score Modeling with HMMs We estimate a patient's future risk of developing transplant rejection by training an HMM to predict the trajectory of risk scores generated by an ANN. We found these risk scores are strong indicators of rejection risk, achieving a mean AUC of 0.92 with 5-fold cross-validation using the rejection labels produced by pathologists as ground truth. An HMM is an interesting model for this task because it allows us to model different sub-clinical phenotypes with the latent states. To find the optimal number of hidden states for our HMM, we calculate two evaluation criteria: (1) the negative log-likelihood of the HMM (calculated using the Forward-Backwards algorithm), and (2) the mean-squared error of our HMM's prediction of the final risk score. A stratified 60/40 train/test split was utilized to train the HMM and obtain all model parameters, and evaluate the models, respectively. HMMs were fit using the open-source `hmmlearn` package [9].

Our experiments show that as the number of hidden states increases (from 2 to 20), the negative log-likelihood of our model decreases monotonically (Figure 2). We anticipate that this monotonic decrease in log-likelihood is an artifact of overfitting, since more complex systems are better at modeling idiosyncrasies in training data. However, they are usually are less generalizeable. This is supported by the MSE, which shows that accuracy on the test set decreases when too many hidden states are applied. In this metric, we find that 9 hidden states produces the lowest error ($\mathrm{MSE} = 0.07$). Although this is a large number of latent states, it may not be inappropriate for a phenotype as complex as immune rejection.

Finally, we repeat this experiment, but now constrain the HMM to have an upper triangular transition matrix. This is equivalent to assuming that immune rejection is irreversible (i.e., in this model, once someone transitions to a more severe state, it is impossible to transition back). We know that this is not a clinical reality: patients can transition from rejection back to non-rejection if their clinician increases the dosage of their immunosuppresant drugs. This experiment serves as comparison to show that our fully connected (ergodic) HMM is able to capture the nuances of immune rejection, as it can model not only worsening progression of immune rejection, but also improvements in disease state (Figure 2).

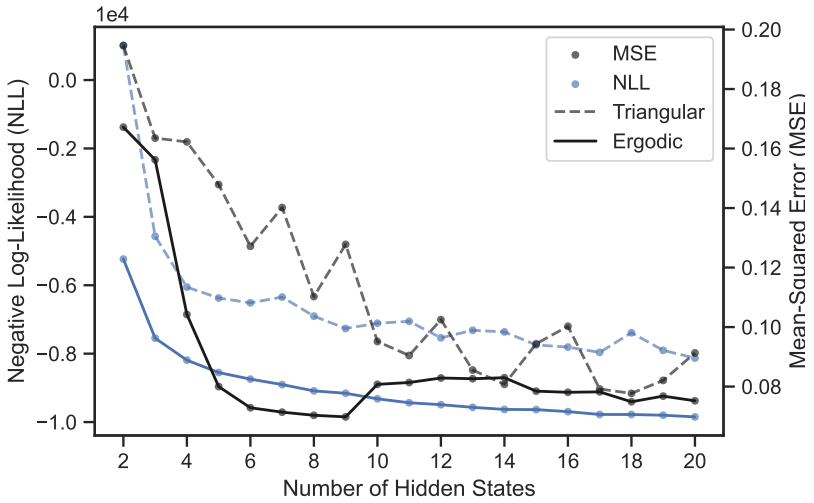

Figure 2: HMM model selection shows that an ergodic HMM with 9 hidden states is the best model with these evaluation metrics over our data set.

3.2 Rejection Classification with LSTMs The LSTM model was tested in five different experiments to understand the effect of longitudinal data on predicting Y_n. these results are summarized in Table 1. As expected, $M(x_{n-1})$ outperforms $M(x_1)$, suggesting that the most recent biopsy is more indicative of future rejection than the first biopsy. Also as expected, model performance improves when longitudinal data is included. Interestingly, $M(x_1, xn_j - 1)$ outperforms $M(x_1, \dots x_{n_j-1})$, which may suggest that our architecture is not sufficiently complex to innately and autonomously capture the trajectory of immune rejection. Finally, providing the full dataset and rejections labels for $M(\hat{x}_1, \dots \hat{x}_{n_1})$ leads to a substantial improvement in performance compared to all other models. We hypothesize that this label serves as a highly valuable feature which assists the model in contextualizing the feature values. It must be noted that for 85% of the patients, $y_{n_j-1} = y_{nj}$, which may explain a large component of the performance increase. Overall, the improvements in performance as the amount of time-series data increases demonstrates that the LSTM is learning from the progression of heart phenotype captured by each successive biopsy. ROC curves are shown in Appendix B

Experiment	Precision	Recall	F1	AUC
x_1	0.62 ± 0.06	0.62 ± 0.04	0.57 ± 0.07	0.57 ± 0.05
x_{n-1}	0.75 ± 0.05	0.72 ± 0.02	0.70 ± 0.03	0.68 ± 0.03
$[x_1, x_{n-1}]$	0.81 ± 0.03	0.80 ± 0.03	0.80 ± 0.04	0.79 ± 0.04
$[x_1, ..., x_{n-1}]$	0.76 ± 0.06	0.75 ± 0.06	0.75 ± 0.06	0.73 ± 0.06
$[x_1, ..., x_{n-1}]$ with labels	$\mathbf{0.94 \pm 0.02}$	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.96 \pm 0.04}$

Table 1: Average metrics ± 1 SD for various experiments ran with the LSTM.

4 Discussion In current medical practice, pathologists examine cardiac biopsies to detect immune rejection only at the present time of biopsy, rarely examining past biopsies. Our models demonstrate the value of autonomously examining a patient's entire biopsy history to estimate the future trajectory of a patient's rejection risk. Our experiments show that analysis of longitudinal data outperforms examination of single biopsies, suggesting that immune rejection is not solely a function of the current histologic appearance, and rather an accumulation of immune activity over time. The high accuracy of our models in estimating future probability of rejection shows promise to improve heart transplant patient outcomes. Our LSTM and HMM systems can identify rejection risk in the near-future, showing the potential to improve patient outcomes by enabling proactive preventative therapies and more frequent followups for patients at high risk of rejection, and decreasing treatment burden for patients at low risk.

One drawback of our current approach is that we remain relatively agnostic to time. For example, for two patients with the same number of biopsies, the corresponding feature vectors are fed into the same layers of our LSTM and HMMs, even though the patients may have received those biopsies at immensely different time points. Intuitively, we would expect that time between biopsies is a substantial factor in immune rejection likelihood. In future work, we intend to include time from transplant as an input into our models, as well as experiment with different imputation schemes which are not agnostic to time (such as imputing with time since biopsy on the x-axis, rather than our current approach of biopsy index on the x-axis). Finally, we hope to later consider covariate adjustment in both of our models (i.e., variables like sex, age, immunocompromised, smoking habits, family history, drug adherence, etc), since future rejection is not simply a function of immune activity through time, but also of these covariate factors. Since WSIs will likely have data set shift because different institutions due to variations imaging protocols, training our models on non-noisy covariates should help make them more robust. Lastly, we have trained and tested on patients all treated and scanned at the same hospital. In the future, we would like to test our models on patients from different institutions to better evaluate the generalizability of our models to different staining techniques, patient demographics, and scanner artifacts.

Author contributions are listed in Appendix F.

Appendix A. Methodology Overview.

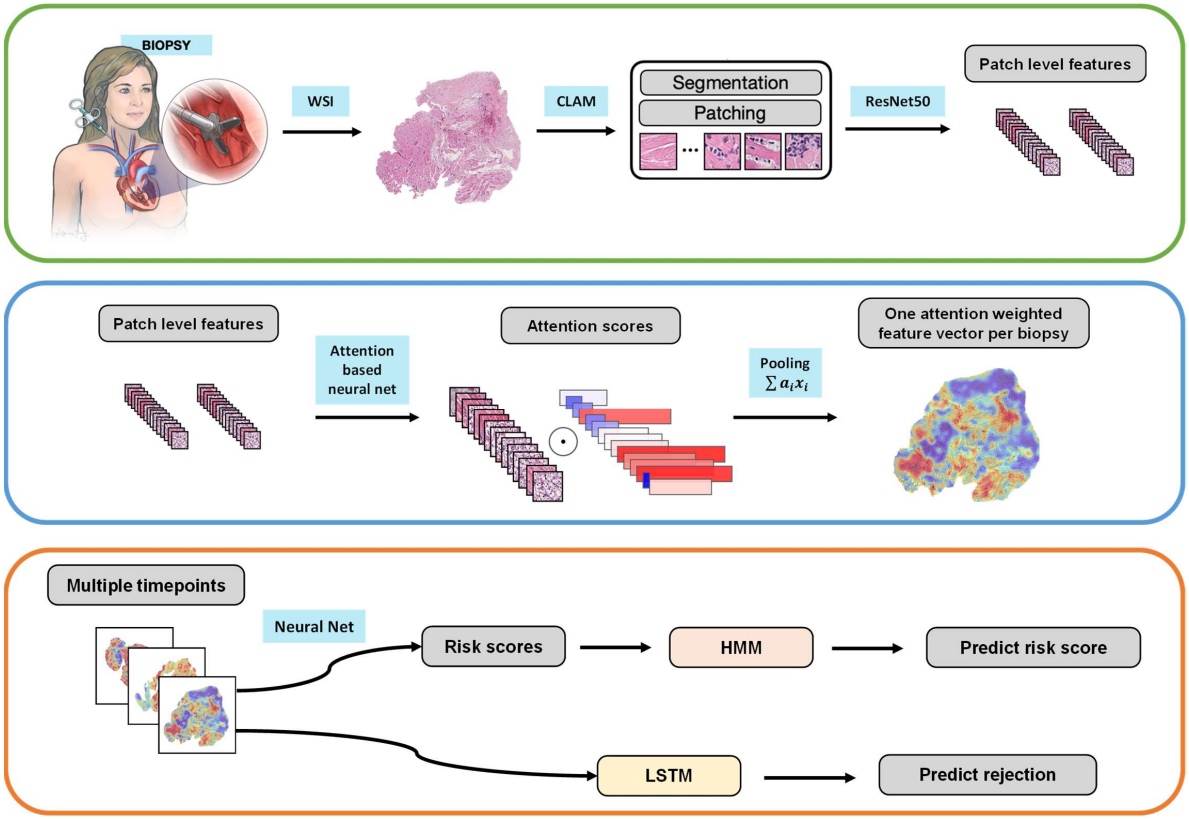

Figure 3: Biopsies are stained and imaged to create Whole Slide Images (WSIs). Features are extracted from these images using a weakly supervised neural network. Finally, the risk of future immune rejection is modeled using statistical modeling (HMM) and neural networks (LSTM).

Appendix B. ROC Curves for LSTM.

Appendix C. Attention Pooling. We apply an attention-based pooling function to identify patches of high diagnostic value. In attention-based pooling, attention scores are learned for each feature vector of a WSI, such that the attention score a_k for the k'th feature vector is computed as follows:

$$a_k = \frac{\exp\{w^T \tanh(Vh_k^T)\}}{\sum_{j=1}^{K} \exp\{w^T \tanh(Vh_j^T)\}\}}$$

Where $w \in \mathbb{R}^L$ and $V \in \mathbb{R}^{L \times M}$ are parameters learned by a neural network. To allow a variable number of feature vectors per biopsy, the scores are normalized such that $\sum_{i=1}^{m} a_i = 1$. This step is critical to the performance of our models due to the nature of immune rejection. Attention based pooling is a type of multiple instance learning (MIL) which fits closely with the requirements of our task of identifying immune rejection. In the MIL framework, every biopsy of a patch is an instance, and the patches from a biopsy constitute a bag. If even a single patch identifies abnormal immune activity, then there is a high likelihood of immune rejection. Consequently, the identification of attention scored through weakly supervised learning strengthens our ability to extract clinically meaningful feature representations.

Attention scores indicating the diagnostic value of each patch are then generated from analysis

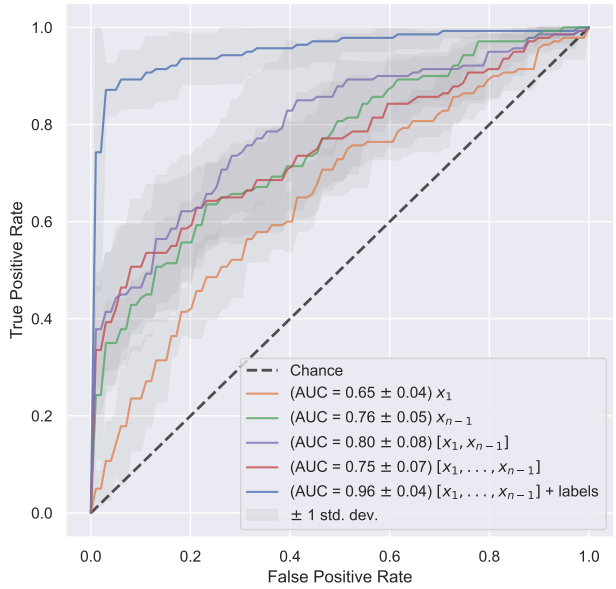

Figure 4: Mean ROC curves with standard deviation calculate dusing the five fold cross validation study for the vrious LSTM experiments.

of the feature vectors and biopsy-level ground truth corresponding to each WSI. In order to generate a single feature vector $\hat{x}$ for each biopsy, we average the extracted feature vectors weighted by their corresponding attention scores.

Appendix D. Generating Risk Scores. We trained an artificial neural network (ANN) to identify immune rejection at the time of biopsy. Formally, for the set of ground truth labels, $Y' = [Y_1, Y_2, \ldots, Y_n]$, the ANN takes the embedding $\hat{x}_{ij}$ of biopsy j from patient i, and predicts $(Y_i')_j$. To learn histology-specific features representations, the deep features are fed through two stacked fully-connected layers L_{c_1} and L_{c_2} with biases and weights $W_1 \in \mathbb{R}^{512\times1024}$, $b_1 \in \mathbb{R}^{512}$ and $W_2 \in \mathbb{R}^{512\times512}$, $b_2 \in \mathbb{R}^{512}$, both with ReLU activation functions. In sum, the weighted ResNet features $\hat{x}_{ij}$ are mapped to a 512-dimensional vector

$$h_{ij} = \mathrm{ReLU}\left(W_2\left(\mathrm{ReLU}\left(W_1 z_{ij}^T + b_1\right)\right) + b_2\right) .$$

Finally, h_{ij} is fed into an output layer parameterized by $W_{out} \in \mathbb{R}^{2\times512}$ and $b_{out} \in \mathbb{R}^2$ and a softmax activation function to output the probability of belonging to each class (no rejection or rejection)

$$p = \mathrm{Softmax}(W_{out} h_{ij} + b_{out}) ,$$

where p is the probability that patient i is experiencing rejection at biopsy j.

We apply the fast-gradient sign method (FGSM) to identify the largest step possible which maximizes loss with respect to perturbation δ. Specifically, we calculate the perturbation δ for each specific input embedding as

$$\delta_{ij} = \epsilon \mathrm{sign}(\nabla_x l(h_0(\hat{x}_{ij}), y_{ij}))$$

We then modify the input embedding to be $x'_{ij} = \hat{x_{ij}} + \delta_{ij}$.

Appendix E. Interpolation Methodology. Each patient in the dataset has a different number of biopsies, but the methods explored in this study require the same number of time points for each observation. In natural language processing, an analogous problem (modeling sentences with different

number of words) is solved by adding 0-padding [10]. This simplistic solution does not work with modeling sequences of images. Thus, the following strategy was used to ensure same number of data points per patient:

- Find the maximum number of biopsies per patient in the dataset
- For each patient, uniformly space out their biopsies over the maximum number of biopsies. Use zero vectors as a placeholder for missing biopsies
- Using linear interpolation, generate biopsies between two time points with known biopsy features. Repeat this step to complete the dataset
- In order to interpolate the labels, pull down interpolation was used (last known label is carried forward and assigned to interpolated data until the next known label is encountered)

The same linear interpolation strategy is used for risk scores. While linear interpolation allows rapid imputation of missing data, more intricate generative modeling methods can be explored in the future for data imputation.

Appendix F. Contributions. All team members contributed equally to the writing of deliverables. With regards to computational division of labor:

- Daniel Shao: Feature extraction and risk score generation
- Vivek Gopalakrishnan: HMM experiments
- Anurag Vaidya: LSTM experiments

References.

[1] Matthew J Everly. Cardiac transplantation in the united states: an analysis of the unos registry. *Clinical transplants*, pages 35–43, 2008.

[2] J G Shanes, J Ghali, M E Billingham, V J Ferrans, J J Fenoglio, W D Edwards, C C Tsai, J E Saffitz, J Isner, and S Furner. Interobserver variability in the pathologic interpretation of endomyocardial biopsy results. *Circulation*, 75(2):401–405, February 1987. doi: 10.1161/01.CIR.75.2.401.

[3] Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170–175, October 2016. ISSN 1361-8415. doi: 10.1016/j.media.2016.06.037.

[4] Faisal Mahmood, Daniel Borders, Richard J. Chen, Gregory N. Mckay, Kevan J. Salimian, Alexander Baras, and Nicholas J. Durr. Deep Adversarial Training for Multi-Organ Nuclei Segmentation in Histopathology Images. *IEEE transactions on medical imaging*, 39(11):3257–3267, November 2020. ISSN 1558-254X. doi: 10.1109/TMI.2019.2927182.

[5] Chengkuan Chen, Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Andrew J. Schaumberg, and Faisal Mahmood. Fast and Scalable Image Search For Histology. *arXiv:2107.13587 [cs, q-bio]*, July 2021.

[6] Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, June 2021. ISSN 2157-846X. doi: 10.1038/s41551-020-00682-w.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[8] Rafid Sukkar, Elyse Katz, Yanwei Zhang, David Raunig, and Bradley T. Wyman. Disease progression modeling using Hidden Markov Models. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2012:2845–2848, 2012. ISSN 2694-0604. doi: 10.1109/EMBC.2012.6346556.

[9] Hmmlearn/hmmlearn. hmmlearn, December 2021.

[10] Mahidhar Dwarampudi and N. V. Subba Reddy. Effects of padding on LSTMs and CNNs. *arXiv:1903.07288 [cs, stat]*, March 2019.