Using Deep Learning to Diagnose Juvenile Idiopathic Arthritis in the Temporomandibular Joint

Eric Loreaux [1] **Nicolas Bievre** [2] **Kaylie Zhu** [3] **Huaiyang Zhong** [4]

Abstract

In this study, we investigate the application of deep learning models to the visual diagnosis of juvenile idiopathic arthritis present in the temporomandibular joint (TMJ). The primary challenge is the size of our dataset (100 patients). We test two different methodologies, first using each patient scan as a training example and training end-to-end diagnoses, and second training models on single MRI slices within a patient scan as a technique for boosting the amount of training data. The outputs of these models are at the slice level, and so we combine outputs for all slices of a given patient using a custom voting scheme. Our results show that this slice-level methodology is most effective at predicting the presence of arthritis in the TMJ, with an AUC score of 1.0 and an F1 score of 0.952 on the validation set.

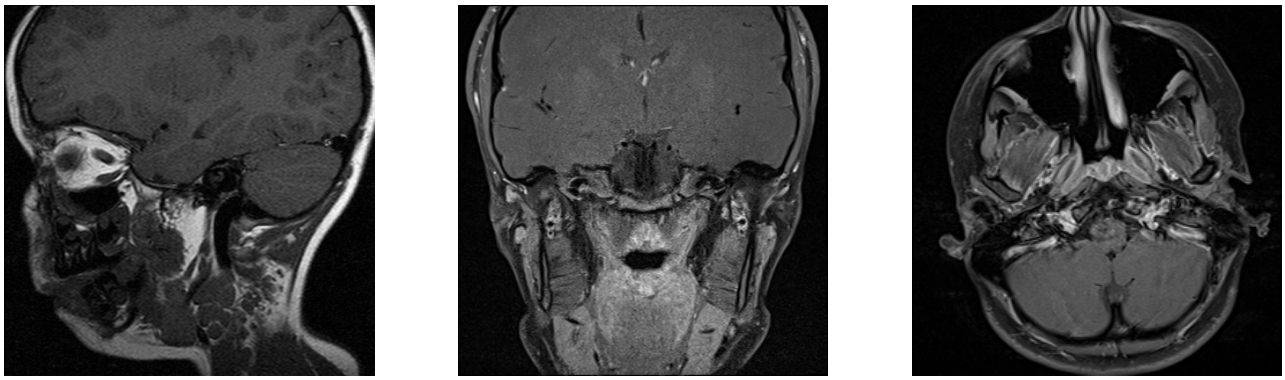
1. Introduction

Juvenile Idiopathic Arthritis (JIA) is a chronic inflammatory disease that affects 1 in 1000 children in the United Kingdom (Arvidsson et al., 2009). The disease can manifest in the jaw area - more precisely in the Temporomandibular Joint (TMJ) and may entail facial growth disturbances, pain, and/or impaired jaw function (Larheim et al., 2015). Magnetic Resonance Imaging (MRI) remains the technique of choice for radiologists looking to monitor arthritis in the TMJ (Arvidsson et al., 2010), (Navallas et al., 2017) and determine whether intervention is necessary. This intervention entails a potentially painful steroid injection into the TMJ space (Stoustrup P, 2015). The problem with using MR imaging for diagnosis is the large degree of variability amongst radiologists, which can undermine the objective truth of the disease state in patients and put some at risk for unnecessary intervention.

There is a need for a more robust method of quantifying arthritis that is not subject to the biases of individual assessors and leverages the current data available for these patients. The effect of such a method would be to introduce consistency into the diagnosis of TMJ arthritis and to reduce the number of false positives that may lead to unnecessary, invasive, and painful treatments. Paediatric Rheumatologists at The Bristol Royal Hospital entered a collaboration with Stanford University via the project class CS 341 to develop such a quantification method. The goal of this project is to investigate the application of deep neural network models to a dataset of MRI scans of both healthy patients and those afflicted with JIA, with the goal of determining the effectiveness of these models for consistent diagnosis of TMJ arthritis.

In this study, we report results for two separate approaches to the problem of accurate TMJ arthritis diagnosis. In the first approach, we implement a model similar to MRNet (Bien et al., 2018), which treats each patient scan as a single training example and outputs a single diagnosis. MRNet was initially developed and validated on a dataset containing 1,370 MRI exams of the knee, however in this case, our dataset is smaller by a factor of ten. We are interested in seeing how such an architecture extends to datasets of this size. In the second approach, we propose a novel method best suited for a severely limited amount of data. This method involves treating every individual slice in each MRI scan as a single training example, which expands our training set significantly. For each slice, we generate two outputs: the first relates to the diagnostic relevance of the slice - many slices do not contain the joint of interest - and the second relates to the actual diagnosis. We explore different voting schemes to combine model outputs for individual slices into a single diagnosis.

[1]Stanford University Department of Bioengineering [2]Stanford University Department of Statistics [3]Stanford University Department of Computer Science [4]Stanford University Department of Management Science and Engineering. Correspondence to: Eric Loreaux <eloreaux@stanford.edu>, Nicolas Bievre <nbievre@stanford.edu>, Kaylie Zhu <kayliez@cs.stanford.edu>, Huaiyang Zhong < hzhong34@stanford.edu>.



(a) Sagittal orientation (b) Coronal orientation (c) Axial Orientation

Figure 1. Magnetic Resonance Imaging of the head of a patient with the 3 different orientations.

2. Related Work

In recent years, deep learning has increasingly become a powerful method for modeling the complex relationships between medical images and their interpretations. Recent advancements in deep learning and large datasets have enabled algorithms to outperform medical professionals in a wide variety of medical imaging and diagnostics tasks, including diabetic retinopathy detection (Gulshan et al., 2016), skin cancer classification (Esteva et al., 2017), arrhythmia detection (Rajpurkar et al., 2017a), hemorrhage identification (Grewal et al., 2017) as well as pneumonia (Rajpurkar et al., 2017b). While not capable of medical autonomy, these models are proving to be useful diagnostic decision support tools in a variety of medical contexts (Sayres, 2018).

Magnetic resonance imaging (MRI) is a common method for diagnosing various types of pathologies. Automated diagnosis from MRI scans has received growing attention across the medical industry, with applications ranging from knee tear assessment to schizophrenia (Zeng et al., 2018). Islam et al. (2017) studied the performance of various convolutional architectures on Alzheimer's disease diagnosis using the publicly available Open Access Series of Imaging Studies dataset (Demner-Fushman et al., 2015). More recently, (Bien et al., 2018) developed an automated machine-learning model for detecting knee abnormalities and measured the clinical utility of providing model predictions to clinical experts during interpretation. In this study, the deep learning system is leveraged to decrease diagnostic error and variability, as well as improve efficiency of specific diagnoses (anterior cruciate ligament [ACL] tears and meniscal tears) on knee MRI exams.

3. Dataset

3.1. Overview

For this project, we start with a dataset of MRI scans for 151 patients. Of these patients, 53 do not have JIA, whereas 98 present with JIA symptoms in the Temporomandibular Joint (53N,98Y). Each patient has between 3 and 16 scans, each of which is comprised of a series of slices through the head region with the same spatial orientation and MRI protocol. There are 131 unique protocols contained in the dataset, and many of these map to familiar MRI protocols such as T1, T1 fs Gd, and T2 fs. There are three unique spatial orientations, as seen in **Figure 1**: sagittal (side to side), axial (top to bottom), and coronal (front to back). This dataset was provided by collaborators at the Bristol Royal Hospital. Ground-truth labels show the existence of arthritis and the degree of severity (mild, moderate, severe) for each side of the jaw. These labels were determined empirically after further clinical investigation. Most patients possess scans from multiple dates, as MRI is used to monitor the effectiveness of treatment and the state of disease progression.

3.2. Filtering

For this analysis, we use scans taken on the earliest date for each patient. This way, we ensure that no intervention has yet been administered to the patient. We also select only scans in the axial orientation. Taken along the vertical axis of each patient, these scans contain slices which include both left and right TMJ regions simultaneously. With help from collaborators in Bristol, we identify 37 MRI protocols that all map to the standardized protocol T1 fs Gd. This protocol, which relates to a T1 MRI scan taken after the patient is administered with a contrast agent (Gadolinium), has been shown to be most diagnosti-

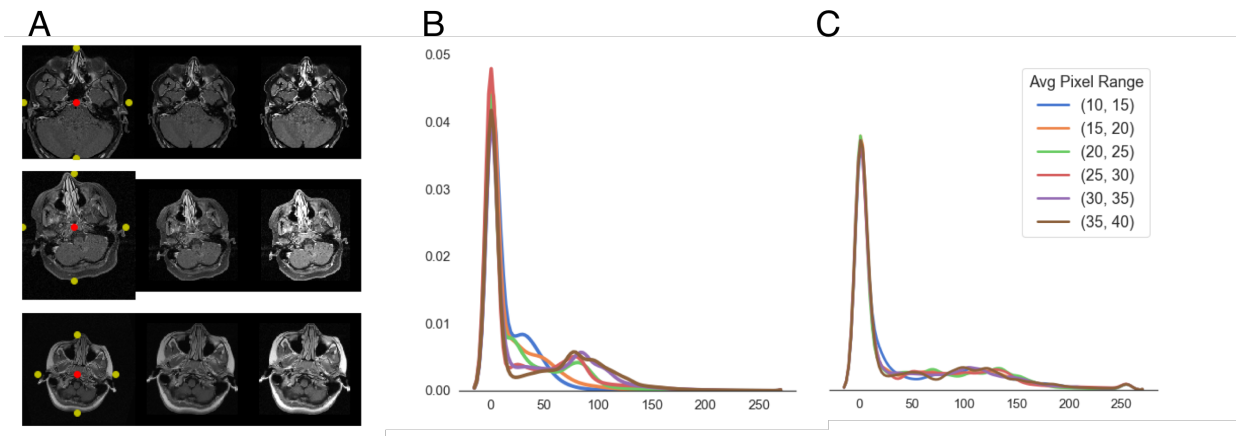


Figure 2. Examples from the data preprocessing pipeline. **A** Images vary in shape, enlargement, and intensity, and so they are centered using the skull as a reference, cropped to a square shape, and normalized via histogram normalization. **B,C** Intensity histograms before and after normalization. Six representative densities are displayed from varying average pixel ranges. Before normalization (B), these bimodal densities are not aligned, and after normalization (C), they have been brought into similar ranges.

cally relevant in the detection of TMJ arthritis (Kellenberger et al., 2018). Ultimately, these filtering decisions, along with further filtering to remove low quality scans with movement artifacts, reduce our dataset to 114 patients (38N/76Y). Each patient has between 1 and 3 scans, and each scan is comprised of between 9 and 144 slices ($\mu = 28.2, \sigma = 28.7$), which leads to a final dataset of $4,222$ slices.

3.3. Preprocessing

When comparing MRI images across different patients, we found that the shape of the image, the location of the patient's head, and the degree of enlargement of the head can vary significantly. One of the additional challenges with MR images is that the actual intensity scales do not have a fixed meaning and can vary between scans. To mitigate these issues of data heterogeneity, proper data preprocessing is critical. All images are centered using the boundaries of the skull as reference points (detected via intensity thresholding) and re-sized to a consistent size of 256x256 pixels (**Figure 2**). To normalize image intensity, we use a biologically relevant normalization approach that involves adjusting the intensity histograms of each image to match the percentiles of a training subset (Nyu & Udupa, 1999). After this normalization technique, all image intensities are scaled and shifted to match the average mean and standard deviation of the training images ($\mu = 45.6, \sigma = 62.8$).

At train time, images fed through the data loader are rotated and shifted randomly, and flipped horizontally with 50% probability. Images are also scaled between 0.7 and 1.3 of the original size. The output of this procedure is a 224 × 224-pixel image, which is then fed as input to the models.

3.4. Training, validation, test

The patients and their corresponding images are split into training, validation and test cohorts with the ratio 80:10:10. Cohorts are split such that the scans and slices related to a single patient are not found in more than one cohort. We use stratified random sampling to ensure that at least 4 positive patient labels of each arthritis severity (mild, moderate, and severe) are present in the validation and test cohorts. The patient counts for these splits can be seen in **Table 1** below.

Cohort	Normal	Abnormal
Train	30	59 (7mild, 39mod, 10sev, 11n)
Val	4	8 (4 mild, 4 mod, 4 sev)
Test	4	9 (4 mild, 4 mod, 4 sev)

Table 1. Train/val/test cohort splits with patient counts. Note that counts of mild, moderate, and severe labels refer to patients with positive indications of these labels in at least one side of the head.

4. Models

The two model architectures and their corresponding approaches are laid out in **Figure 3**. In the first approach, each scan is used as a single training sample. The s individual slices within each scan are fed through a feature extractor resembling the early convolutional

layers of AlexNet (Krizhevsky et al., 2012). The output of this feature extractor is a set of numerical features with dimensions $s \times 256 \times 7 \times 7$. Each slice then undergoes global average pooling, which results in a dimensionality of $s \times 256$. All s vectors are then max pooled and fed through a fully connected layer with a final sigmoid activation to obtain a binary diagnosis.

The second approach is a novel approach that we propose as an alternative to scan-level diagnoses. This approach is best suited for prediction and diagnosis tasks in which there are very few scans available to train on. The primary assumption that underlies this approach is that individual slices may contain enough information for reasonably accurate diagnoses. If each slice is then used as a training example, the amount of training data is increased significantly, boosting the robustness and generalizability of the model. This requires us to accomplish two different binary tasks. In the first task, each image is labeled as diagnostically relevant (containing the joint of interest) or irrelevant (containing some other region of the head). The second task is the diagnostic task, in which relevant images are determined to be normal or abnormal in nature.

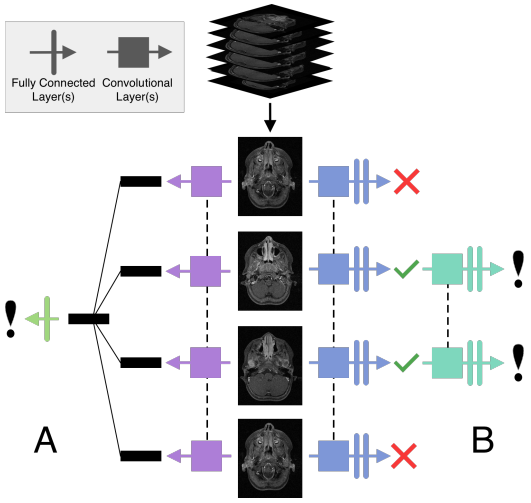


Figure 3. Two different methodologies for generating diagnoses. **A, left** MRNet reads all scan images through a convolutional feature extractor before combining the resultant activations and passing them through a fully connected layer. **B, right** Our approach treats each slice as an individual training example. One model seeks to find slices of diagnostic relevance, and the second model seeks to output slice-level diagnoses.

We also attempt to combine the two models in the second approach into a single multitask model. The input to this model is a single slice, and the output is a softmax probability across three different classes: No jaw, normal jaw, and abnormal jaw. The 'No jaw' label is attributed to diagnostically irrelevant images. This multitask approach allows one model to learn a shared set of weights that accomplish both tasks, which increases regularization and boosts generalizability. It also allows diagnostically irrelevant images to be fed as training data for both tasks, whereas for the split model approach, this data is not available to the diagnostic model.

5. Objective and metrics

For this paper, we investigate the detection of TMJ arthritis as a binary classification problem, in which the features of interest are the pixels of a relevant MRI image, denoted X. Most of our models predict a diagnosis in the form of a binary label $\hat{y} \in \{0, 1\}$. In two of our models, $\hat{y}$ indicates the probability of the presence of the disease, and for our stage I. model in the 2-stage approach, $\hat{y}$ indicates the probability that a scan is diagnostically relevant. For this problem, we use the binary cross entropy loss function to optimize our model:

$$L(X, y) = -y \log p(\hat{y} = 1|X) - (1 - y) \log p(\hat{y} = 0|X)$$

Where y is the true label for each image, determined empirically through clinical assessment. We also explore the possibility of combining two different binary tasks that operate on the same dataset into a single model. In order to train this model, we reformulate the problem as one of multi-class classification by creating a three-label scheme: "no jaw," "normal jaw," and "abnormal jaw". This will require us to use the categorical cross entropy loss function:

$$L(X, y) = -\sum_{c=1}^{4} y_c \log P(\hat{y} = c|X)$$

Where y_c is a one-hot vector with a 1 only at the location of the ground-truth class and $P(\hat{y} = c|X)$ is computed using a softmax transformation of the final activation layer of the model.

Along with accuracy, we will also be tracking F1-score as a key metric of performance for this task. The F-score is a metric that incorporates both precision and recall and can be tuned using the parameter β:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

Task	Architecture	Val Acc	Val F1	Val AUC
Patient Diagnosis	MRNet	0.813	0.870	**0.836**
Patient Diagnosis	MRNet	**0.875**	**0.917**	0.791
I. Slice Relevancy	ResNet18	0.876	**0.797**	**0.954**
I. Slice Relevancy	ResNet34	**0.894**	0.794	0.931
II. Slice Diagnosis	ResNet34	0.797	0.876	**0.806**
II. Slice Diagnosis	ResNet34	**0.872**	**0.925**	0.754
Patient Diagnosis	(I & II) Max Vote	**0.813**	**0.880**	**0.836**
Multitask Slice	PNASNet-5-Large	**0.84**	**0.745**	**0.797**
Patient Diagnosis	Multitask Max Vote	**.938**	**0.952**	**1.0**

Table 2. Performances across all models and voting schemes.

6. Results

All of the best results of all our models can be seen in **Table 2**.

6.1. MRNet

For MRNet, we combine the AlexNet feature vectors for all slices within a single patient scan using vector-level average pooling followed by a global max pool. This combined feature vector is sent through a single fully connected layer, and the output is a single number that is then fed through a sigmoid activation to obtain a binary prediction of disease/no disease. For our feature extractor, we pretrain the convolutional layers of AlexNet on ImageNet (Deng et al., 2009), a repository of millions of images, however we continue to finetune these layers throughout the training process on our smaller dataset.

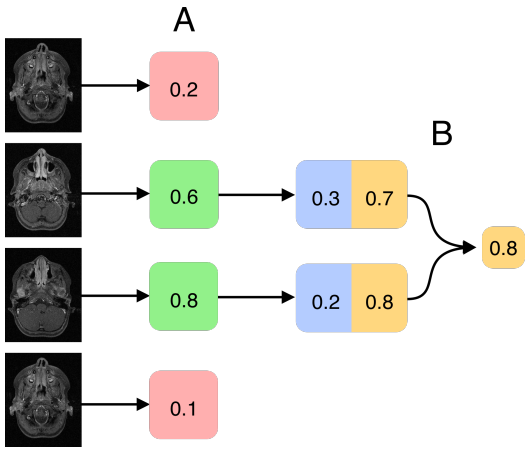


Figure 4. A visual depiction of the general voting scheme. **A** in step 1, single slices are selected based on diagnostic relevancy from stage I. **B** in step two, the diagnosis votes from stage II are combined by taking the highest probability across all outputs.

6.2. SliceNet

For slice-level tasks, we train three different models. The first model, referred to as the stage I model, which is responsible for flagging diagnostically relevant slices (slices that actually depict the jaw and not other regions of the head). The second model is referred to as the stage II model, and this is a model trained to output an accurate diagnosis at the slice-level. The third and final model is a combination of both of the previous models. This model learns a multitask approach, in which the output is the probability that each slice is either 'no jaw,' 'normal jaw,' or 'abnormal jaw'. This output is computed by performing a softmax over the three possible classes.

We found that the task of selecting diagnostically relevant images was easier to learn than that of diagnosis, and we consistently achieved higher validation metrics for this task. We found that the multitask model performed worse than the other two models, likely because the combined labels make the task more difficult. We experimented with a variety of deep convolutional neural network model architectures, including ResNet18, ResNet 34 (He et al., 2015) and PNASNet-5-Large (Liu et al., 2017).

6.3. Voting Scheme

Our primary method for generating patient diagnoses with the slice-level models is to combine their individual outputs across all patient slices according to some voting scheme. The method we implement is a selection by diagnostic relevancy followed by a combination of diagnosis probabilities. We choose a probability threshold past which we assume an image is diagnostically relevant. The voting scheme we devise for all diagnostically relevant images is to take the maximum of all softmaxed probabilities. An example of this maximum voting scheme can be seen in **Figure 4**.

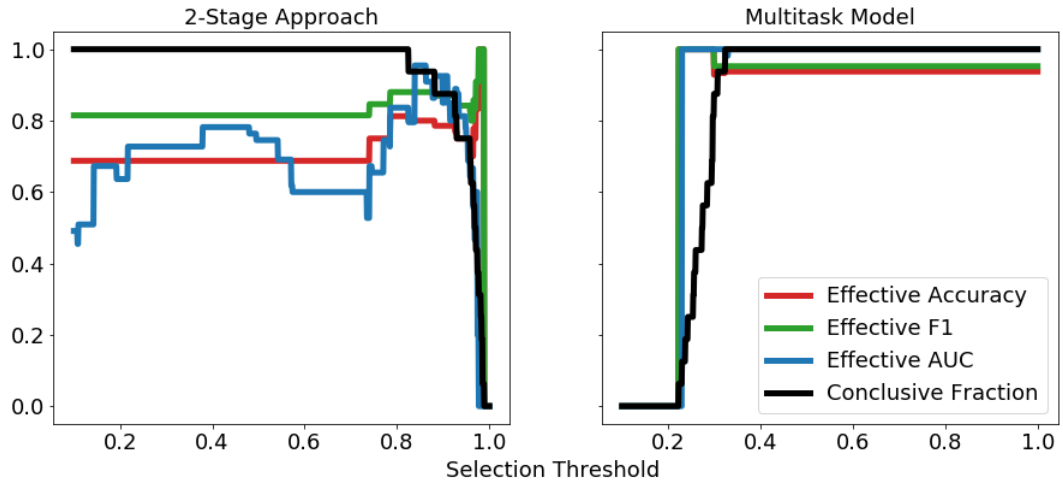


Figure 5. Performances from slice-level model voting for both the 2-stage approach and using the multitask model. The performances are reversed because the selection threshold is used as a minimum threshold in the 2-stage approach and a maximum threshold for the multitask model. The conclusive fraction refers to the fraction of patients for which a conclusive diagnosis can be made. It is impossible to make a diagnosis if no scans are selected in step 1.

We carry out this patient-level voting using both our two-stage slice-level models and our multitask slice-level model. We tune the selection threshold in order to feed different numbers of images to the second step for diagnosis and see how our performances are affected. The performance of these models can be seen in **Figure 5**. For these tasks, the selection threshold is used in opposite ways: For the two-stage approach, the output of the stage I model is the probability that the slice is diagnostically relevant, and so we only use slices with a relevance probability above the selection threshold. For the multitask model, one possible class is 'No jaw,' and the higher this probability, the less likely the image is diagnostically relevant. Therefore, we only use images with a 'No jaw' probability below the selection threshold.

Test metrics have not yet been calculated for these tasks, since model tuning is not finished. We suspect that the slice-level models will ultimately prove more generalizable to the test set than the scan-level MRNet since the expansion of all scans into a multitude of slices forces our models to identify disease signals with much fewer bits of information. Combining these votes using a voting scheme also provides an opportunity for an ensemble-like output for each patient that should be more robust to patient variation.

7. Conclusion

In this project, we set out to investigate the use of neural network models for the visual diagnosis of juvenile idiopathic arthritis in the temporomandibular joint. We chose to use MRI scans of an axial orientation taken with the T1 fs Gd protocol. Due to the small number of patients, we were interested to see whether the leading neural network architecture for MRI-based diagnosis would generalize well. We also devised a new method that involves using each individual MRI slice within a scan as a training example. We then combined the slice-level votes for a single patient scan to obtain a final diagnosis. If effective, this method would prove useful in providing robust decision support tools for medical imaging scenarios in which the datasets are relatively small (rare diseases, limited data access, etc).

Our preliminary results are promising. MRNet performed at the scan level achieves a maximum F1 score of 0.917 and a maximum AUC score of 0.836, but not simultaneously. Our 2-stage model approach (stage I selects diagnostically relevant slices, stage II makes slice-level diagnoses) shows that when votes are combined, we achieve an equivalent AUC score on the validation set to MRNet, the state of the art scan-level model. In addition, when slice-level votes from the multitask model are combined, we outperform all previous approaches with an AUC score of 1.0 and a validation F1 score of 0.952.

8. Future Work

There is a large amount of work left to be done to explore this diagnostic application. Investigating loss

functions that better reflect the metrics of interest for this task (Eban et al., 2017) would be interesting, since false positives and false negatives can mean different things in a medical context. We also need to perform further tuning of voting schemes and their corresponding parameters to identify the best methods for combining slice-level votes. Our ground truth dataset also includes more complex disease labels, such as degree of severity of arthritis (mild, moderate, severe) and labels specific to the left and right sides of the head.

In the future, it would be valuable to incorporate MRI scans of multiple orientations and protocols. We use scans of the axial orientation for this study, however doctors frequently use scans of the sagittal orientation for best results. This orientation, which depicts a sideways view of the head, would contain much more of the TMJ region in a single slice, and so our slice-level models may perform better. Other MRI protocols such as T2 fs have also proven useful for detecting arthritis in the TMJ.

Finally, we would love to continue working with doctors at the Bristol Royal Hospital to develop more robust evidence of the efficacy of these sorts of neural network models as diagnostic support systems. In order to quantify the clinical effect of these systems, an in-depth analysis is required in which the accuracy of clinical diagnosis by expert radiologists is measured with and without the aid of model outputs.

References

Arvidsson, Linda Z., Flat, Berit, and Larheim, Tore A. Radiographic tmj abnormalities in patients with juvenile idiopathic arthritis followed for 27 years. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, 108(1):114 – 123, 2009. ISSN 1079-2104. doi: https://doi.org/10.1016/j.tripleo.2009.03.012. URL http://www.sciencedirect.com/science/article/pii/S1079210409001619.

Arvidsson, Linda Z., Smith, Hans-Jrgen, Flat, Berit, and Larheim, Tore A. Temporomandibular joint findings in adults with long-standing juvenile idiopathic arthritis: Ct and mr imaging assessment. *Radiology*, 256(1):191–200, 2010. doi: 10.1148/radiol.10091810. URL https://doi.org/10.1148/radiol.10091810. PMID: 20574096.

Bien, Nicholas, Rajpurkar, Pranav, Ball, Robyn L., Irvin, Jeremy, Park, Allison, Jones, Erik, Bereket, Michael, Patel, Bhavik N., Yeom, Kristen W., Shpanskaya, Katie, Halabi, Safwan, Zucker, Evan, Fanton, Gary, Amanatullah, Derek F., Beaulieu, Christopher F., Riley, Geoffrey M., Stewart, Russell J., Blankenberg, Francis G., Larson, David B., Jones, Ricky H., Langlotz, Curtis P., Ng, Andrew Y., and Lungren, Matthew P. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of mrnet. *PLOS Medicine*, 15(11):1–19, 11 2018. doi: 10.1371/journal.pmed.1002699. URL https://doi.org/10.1371/journal.pmed.1002699.

Demner-Fushman, Dina, Kohli, Marc D, Rosenman, Marc B, Shooshan, Sonya E, Rodriguez, Laritza, Antani, Sameer, Thoma, George R, and McDonald, Clement J. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23 (2):304–310, 2015.

Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, and Fei-Fei, Li. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. IEEE, 2009.

Eban, Elad, Schain, Mariano, Mackey, Alan, Gordon, Ariel, Saurous, Rif A., and Elidan, Gal. Scalable learning of non-decomposable objectives. *AISTATS*, 2017.

Esteva, Andre, Kuprel, Brett, Novoa, Roberto A, Ko, Justin, Swetter, Susan M, Blau, Helen M, and Thrun, Sebastian. Dermatologist-level classification

of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

Grewal, Monika, Srivastava, Muktabh Mayank, Kumar, Pulkit, and Varadarajan, Srikrishna. Radnet: Radiologist level accuracy using deep learning for hemorrhage detection in ct scans. *arXiv preprint arXiv:1710.04934*, 2017.

Gulshan, Varun, Peng, Lily, Coram, Marc, Stumpe, Martin C, Wu, Derek, Narayanaswamy, Arunachalam, Venugopalan, Subhashini, Widner, Kasumi, Madams, Tom, Cuadros, Jorge, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Islam, Mohammad Tariqul, Aowal, Md Abdul, Minhaz, Ahmed Tahseen, and Ashraf, Khalid. Abnormality detection and localization in chest x-rays using deep convolutional neural networks. *arXiv preprint arXiv:1705.09850*, 2017.

Kellenberger, Christian J., Junhasavasdikul, Thitiporn, Tolend, Mirkamal, and Doria, Andrea S. Temporomandibular joint atlas for detection and grading of juvenile idiopathic arthritis involvement by magnetic resonance imaging. *Pediatr Radiol*, 48:411–426, 2018. doi: 10.1007/s00247-017-4000-0.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012.

Larheim, Tore A., Doria, Andrea S., Kirkhus, Eva, Parra, Dimitri A., Kellenberger, Christian J., and Arvidsson, Linda Z. Tmj imaging in jia patientsan overview. *Seminars in Orthodontics*, 21(2):102 – 110, 2015. ISSN 1073-8746. doi: https://doi.org/10.1053/j.sodo.2015.02.006. URL http://www.sciencedirect.com/science/article/pii/S1073874615000146.

Liu, Chenxi, Zoph, Barret, Shlens, Jonathon, Hua, Wei, Li, Li-Jia, Fei-Fei, Li, Yuille, Alan L., Huang, Jonathan, and Murphy, Kevin. Progressive neural architecture search. *CoRR*, abs/1712.00559, 2017. URL http://arxiv.org/abs/1712.00559.

Navallas, Mara, Inarejos, Emilio J., Iglesias, Estbaliz, Cho Lee, Gui Youn, Rodrguez, Natalia, and Antn, Jordi. Mr imaging of the temporomandibular joint in juvenile idiopathic arthritis: Technique and findings. *RadioGraphics*, 37(2):595–612, 2017. doi: 10.1148/rg.2017160078. URL https://doi.org/10.1148/rg.2017160078. PMID: 28287946.

Nyu, Laszlo G. and Udupa, Jayaram K. On standardizing the mr image intensity scale. *Magnetic Resonance in Medicine*, 42:1072–1081, 1999. doi: https://doi.org/10.1002/(sici)1522-2594(199912)42:6\%3c1072:aid-mrm11\%3e3.0.co;2-m.

Rajpurkar, Pranav, Hannun, Awni Y, Haghpanahi, Masoumeh, Bourn, Codie, and Ng, Andrew Y. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836*, 2017a.

Rajpurkar, Pranav, Irvin, Jeremy, Zhu, Kaylie, Yang, Brandon, Mehta, Hershel, Duan, Tony, Ding, Daisy, Bagul, Aarti, Langlotz, Curtis, Shpanskaya, Katie, Lungren, Matthew P., and Ng, Andrew Y. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR*, abs/1711.05225, 2017b. URL http://arxiv.org/abs/1711.05225.

Sayres, Rory. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *American Academy of Ophthalmology*, 2018. doi: https://doi.org/10.1016/j.ophtha.2018.11.016.

Stoustrup P, et al. Temporomandibular joint steroid injections in patients with juvenile idiopathic arthritis: an observational pilot study on the long-term effect on signs and symptoms. *Pediatr Rheumatol Online J.*, 13(62), 2015. doi: 10.1186/s12969-015-0060-6. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4687278/.

Zeng, Ling-Li, Wang, Huaning, Hu, Panpan, Yang, Bo, Pu, Weidan, Shen, Hui, Chen, Xingui, Liu, Zhening, Yin, Hong, Tan, Qingrong, Wang, Kai, and Hu, Dewen. Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity mri. *EBioMedicine*, 30:74 – 85, 2018. ISSN 2352-3964. doi: https://doi.org/10.1016/j.ebiom.2018.03.017. URL http://www.sciencedirect.com/science/article/pii/S2352396418301014.