Fall 12-14-2023

Utilizing Multitask Transfer Learning for Sonographic Rheumatoid Arthritis Synovitis Grading

Jordan Marie Claire Sanders
Embry-Riddle Aeronautical University, sandej34@my.erau.edu

Utilizing Multitask Transfer Learning for Sonographic Rheumatoid Arthritis Synovitis Grading

Jordan Sanders

Master’s Thesis
Data Science Degree Program
Department of Mathematics
Embry-Riddle Aeronautical University
Daytona Beach, FL

2023

The thesis of Jordan Sanders was reviewed by the following:

Prashant Shekhar
Assistant Professor of Data Science at ERAU
Thesis Advisor, Chair of Committee

Gurjit Kaeley
Professor & Chief, Division of Rheumatology, UF Health
Committee Member

Timothy Smith
Professor and Program Coordinator for Data Science at ERAU
Committee Member

Acknowledgements

I would like to thank Dr. Prashant Shekhar for guiding me through this research as my thesis mentor. Additionally, I would like to thank Dr. Gurjit Kaeley and UF Health for providing the data used in this research, as well as assisting in background knowledge. Thank you to Dr. Timothy Smith for assisting in my progress through the MS Data Science program and for his assistance as a member of my thesis committee. Thank you to everyone else who encouraged and inspired me along this path.

Abstract

Classifying the four sonographic Rheumatoid Arthritis (RA) synovitis grades (Grade 0, Grade 1, Grade 2, and Grade 3) is a difficult problem due to the complexity of the relevant markers. Therefore, the current research proposes a Multitask Transfer Learning (MTL) framework for sonographic RA synovitis grading of Ultrasound (US) images in Brightness mode (B-Mode) and Power Doppler mode.

In the medical community, the lack of reliability of scoring these images has been an issue and reason for concern for doctors and other medical practitioners. The human/machine variability across the acquisition procedure of these US images creates an additional challenge that restricts the development of an efficient automated scoring system. The literature reports the lack of coherency among the doctors' opinions about the grade of arthritis for patients in controlled trials. Motivated by these reasons, the current work moves away from the traditional wisdom of separately scoring B-mode and Power Doppler mode images and poses an MTL framework that jointly learns the features for images across both modes, leading to a more robust automated classifier. The multitask nature of the model also provides additional benefits such as better generalization to blinded test data from the inherent regularization and the ability of the model to be trained by a combination of B-Mode and Power Doppler US images, thereby efficiently handling data scarcity.

Results show the superior performance of the proposed approach compared to traditional machine learning algorithms as well as other standard deep learning models such as Convolutional Neural Networks and Vision Transformers. The mean testing accuracy of our proposed MTL model on B-Mode was 51.55% and on Power Doppler was 61.18%. However, since the boundary between classes is not always clear or defined in RA synovitis grades, the Top 2 success rates have been regarded as another measure of performance in this domain. Accordingly, with a mean B-Mode Top 2 success rate of 80.52% and a Power Doppler Top 2 success rate of 82.50%, the proposed approach can reach a near-human doctor-level classification performance, establishing the usability of this approach.

Contents

Chapter 1

Introduction

1.1 Problem Statement and Objectives

In recent times, data science has been gaining in popularity partially due to the advancement in technology and growing amounts of data. This is especially true in the medical world Blaivas et al. (2020); Wu et al. (2022); McMaster et al. (2022). For example, Natural Language Processing (NLP) has been used to classify temporal artery biopsy reports. Deep learning was used to predict future diagnoses and disease activity with time-sequence data from Electronic Health Records (EHRs). Furthermore, Convolutional Neural Networks (CNN) have been used for many medical applications, including determining if joints are healthy or unhealthy McMaster et al. (2022).

This research aims to use deep learning algorithms to perform Rheumatoid Arthritis (RA) synovitis grade scoring on Power Doppler Ultrasound (US), referred to as Power Doppler mode throughout, and Gray-scale US images, referred to as B-Mode (also known as Brightness Mode) throughout. These images contain the dorsal Metacarpophalangeal (MCP) Joint 1 to 5.

There are four grades of RA with synovitis grading: Grade 0, Grade 1, Grade 2, and Grade 3. Since there are a discrete number of classes as the desired predicted value, classification models are used for RA synovitis grade prediction. Computer vision models are used in this research to assist with classification or to classify themselves since the input data is in image format. Multiple statistics are evaluated with the models to evaluate and compare their effectiveness. The ultrasound images are split into three datasets: training, validation, and testing. The training dataset initializes the model. The validation dataset tunes the initial model created with the training set to avoid any potential overfitting (learning the training data's noise) issues happening with the initial model. Lastly, the final reporting of the model is based on its performance with the testing set (a dataset that contains no data that the model viewed during the training or tuning phases). This reporting gives a census on how the model would react to unseen data.

The goal of this research is to create a robust automated classifier to improve the reliability of sonographic RA synovitis scoring and to improve the precision of sonographic RA synovitis scoring. This score evaluates the inflammation of the joint. With a reliable and precise evaluation of the state of the inflammation in the joint, a proper treatment plan for the inflammation can be created.

1.2 Background

As previously stated, deep learning has been gaining popularity and use in the medical world Blaivas et al. (2020); Wu et al. (2022); McMaster et al. (2022); Momtazmanesh et al. (2022).

For example, Dr. Blaivas, Dr. Arntfield, and Dr. White utilized deep-learning techniques "... to identify vessels, bones, nerves, and tendons on transverse upper extremity (UE) ultrasound (US) images." Using a YOLOv3 deep learning algorithm, these doctors could accurately identify "...four common structures on cross-sectional US imaging of the UE..." Blaivas et al. (2020). Although this paper focused on the identification of vessels, bones, nerves, and tendons in ultrasound images instead of the classification of a disease, this still relates to the goal of this research by proving deep learning algorithms can be used to interpret ultrasound images for medical purposes.

Another example of using deep learning algorithms in the medical world comes from a paper called *A deep learning classification of metacarpophalangeal joints synovial proliferation in rheumatoid arthritis by ultrasound images* by Dr. Wu et al. This research aimed to determine if a classification model could be used to classify RA in the metacarpophalangeal joint with an ultrasound image. They used a DenseNet-based deep learning model to predict the ultrasounds' Outcome Measures in Rheumatology (OMERACT) European Alliance of Associations for Rheumatology (EULAR) synovitis scores (OESS). This paper concluded that their model could be used as "...an alternative to the initial screening of RA..." and "...become a useful tool in clinical RA diagnosis..." Wu et al. (2022). The goal of the Wu et al. paper aligns closely with the goal of the current research, as the authors of this paper want to use deep learning to predict the RA OESS. However, the deep learning technique differs between Wu et al.'s research and the current research's proposed approach.

Another source described various ways deep learning has been utilized in the medical world. For example, natural language processing (NLP) has been used for classifying temporal artery biopsy reports, predicting future diagnoses with electronic health reports (EHRs), and predicting disease activity with EHRs. Additionally, convolutional neural networks (CNNs) have been used for HEp-2 image classification, synovial classification of ultrasounds, and identification of halo sign-positive or halo sign-negative pixels. McMaster et al. suggest developing a model to learn with limited data, using transfer learning, using self-supervised learning, and utilizing deep learning methods to increase the data set size in their future studies section McMaster et al. (2022). This relates to the current research by discussing different deep learning methods already established in the medical world and the need to use a deep learning method to increase the data set size before classification.

One survey reviewed the current state of the use of artificial intelligence (AI) in the realm of RA and found that Machine Learning models (e.g., Logistic Regression, Random Forest, Support Vector Machine, LASSO Regression, etc.) have been used to assess RA development risk based on factors such as anti-citrullinated protein antibodies, single-nucleotide polymorphisms, and certain eye diseases. Additionally, they discovered that Machine learning and deep-learning models, including DenseNet, Random Forest, Convolutional Neural Networks, and Artificial Neural Networks, have been used on imaging data, such as ultrasounds, X-rays, and MRIs to diagnose RA Momtazmanesh et al. (2022).

1.3 Data Source

The data set used for this project contains Power Doppler US and B-Mode US images of dorsal Metacarpophalangeal (MCP) Joint 1 to 5. These images were obtained during four clinical trials: Musculoskeletal Ultrasound in Predicting Early Dose Titration with Tocilizumab (RASTS) study, Ultrasound Assessment of Rheumatoid Arthritis Patients Who Changed Diet (RAWLUS), Therapeutic Response of Cannabidiol in Rheumatoid Arthritis (CRASE), and Use of Acthar in Rheumatoid Arthritis (RA) Related Flares (ACTHAR). The RASTS study (ClinicalTrials.gov Number NCT01717859) was a collaboration between UF-COMJ and UCLA, approved by UCLA IRB (#12-001547) and WIRB. The goal of this study was to examine the efficacy of tocilizumab on ultrasound scores. The RAWLUS study (ClinicalTrials.gov #NCT02881307) was performed by UCLA and approved by UCLA IRB (#16-000881). The purpose of this study was to evaluate if a weight loss intervention improves rheumatoid arthritis disease activity. Additionally, the CRASE study (Clinical-

Trials.gov Number NCT04911127) was performed by UCLA and approved by the UCLA IRB (#19-001551). CRASE's goal was to examine the efficacy and safety of CBD treatment as adjunctive to the medical management of RA patients. Lastly, ACTHAR (ClinicalTrials.gov Number NCT02541955) was also conducted by UCLA and approved by the UCLA IRB (#15-001199). The purpose of this UCLA research study was to determine whether musculoskeletal ultrasound inflammatory scores and/or disease activity scores improve with Acthar treatment in rheumatoid arthritis flare. All images have a unique non-identifiable study number. Furthermore, no unique patient identifiers were included in the imaging data, and images were binned according to scores. The B-Mode and Power Doppler US images were scored with the semi-quantitative LA-JAX scoring system Ranganath et al. (2022).

All images in these trials were requested from UCLA and required a data-use agreement between UF-COMJ and UCLA. The US images were saved as JPEG, TIFF, or PNG file types. The images were grouped by anatomical region and imaging "slice." These folders were uploaded to UF one drive and shared with ERAU (#23-085) via dropbox.

There are two overlying datasets: one with B-mode settings and one with Power Doppler settings. B-Mode is a type of grey-scale imaging that produces a 2D image "... in which the organs and tissues of interest are depicted as points of variable brightness" Murphy (2023). Power-Doppler, on the other hand, is a type of ultrasound imaging that "... displays the strength of the Doppler signal in color" Babcock et al. (1996). The difference between the two models can be seen in Figures 1.1 and 1.2.

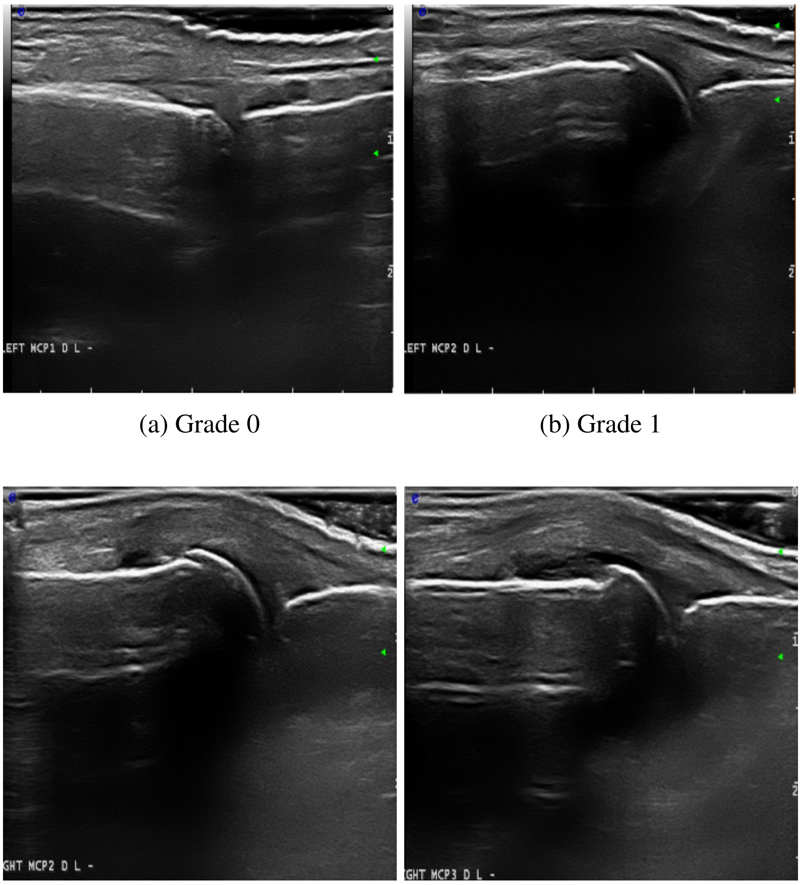
The B-mode setting images have 415 Grade 0 images, 156 Grade 1 images, 465 Grade 2 images, and 205 Grade 3 images. This means 33.44% of the data is Grade 0, 12.57% is Grade 1, 37.47% is Grade 2, and 16.52% is Grade 3 (Table 1.1). Sample images of each grade are shown below in 1.1.

The dataset with the Power Doppler settings contained 612 Grade 0 images, 148 Grade 1 images, 260 Grade 2 images, and 195 Grade 3 images. Therefore, 50.37% was Grade 0, 12.18% was Grade 1, 21.40% was Grade 2, and 16.05% was Grade 3 (Table 1.1). Sample images of each grade are shown below in 1.2.

	Grade 0	Grade 1	Grade 2	Grade 3
B-Mode	415	156	465	205
Power Doppler	612	148	260	195

Table 1.1: Table showing the breakdown of images by sonographic RA synovitis grade and by US image mode.

Each ultrasound mode is split into three different datasets (training, valid-



(a) Grade 0

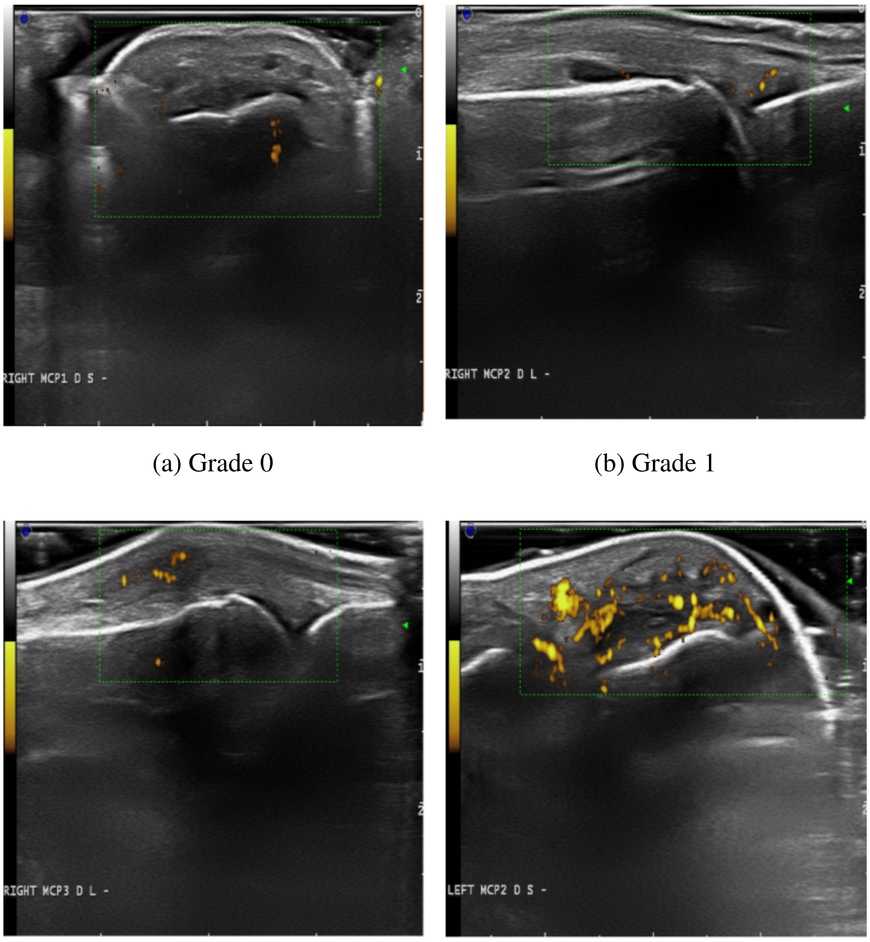
(b) Grade 1

(c) Grade 2

(d) Grade 3

Figure 1.1: B-Mode Ultrasound Image Examples of each RA synovitis grade.

ation, and testing), resulting in six total datasets. The training sets consist of approximately 45% of their respective ultrasound mode's images and serve to be the model's first round of learning. The validation sets contain approximately 30% of their respective US mode's images and are used to tune the models' hyperparameters, which is necessary to avoid overfitting on the training sets. Lastly, the testing sets have approximately 25% of their respective mode's US images. The testing set determines the model's ability to classify unseen data and is, therefore, used to report final accuracies and other evaluation statistics. The exact decomposition of images per dataset is shown in Table 1.2.



(a) Grade 0

(b) Grade 1

(c) Grade 2

(d) Grade 3

Figure 1.2: Power-Doppler Ultrasound Image Examples of each RA synovitis grade.

	Training	Validation	Testing
B-Mode	558	373	310
Power Doppler	546	365	304

Table 1.2: Table showing the breakdown of images by dataset.

Chapter 2

Transfer Learning with Classic Machine-Learning Models

Before jumping into deep neural networks, classic machine learning models were conducted as a baseline. Since these models were not necessarily trained on images, feature extraction was used with ResNet50, which was trained on images. The extracted features were the inputs for the classic machine learning models.

2.1 Feature Extraction with ResNet50

Deep neural networks have been used for image recognition for a while now. However, these models can be extremely complex and computationally expensive. He, Zhang, Ren, and Sun sought to solve this issue by using residual networks (ResNet for short) He et al. (2015). They competed in the ILSVRC 2015 classification task and won first place by achieving a 3.75% testing error on an ensemble of the residual nets.

A plain convolutional network is turned into a residual network by inserting shortcut connections and identity mapping. He et al. found that the residual networks decreased in training error with deeper networks (i.e., more layers), unlike plain networks, meaning the residual networks solve the degradation issues plain networks were experiencing and allow for deeper network use. Furthermore, they determined that residual networks ease optimization and converge faster with comparable accuracy in less deep networks as in plain networks He et al. (2015).

The ResNet paper discovered that their residual network can work for image recognition on multiple datasets. The results discussed thus far have been on the ImageNet dataset, which consists of over a million training images, 50,000 validation images, and 100,000 testing images belonging to 1,000 classes. As previously mentioned, an ensemble of these ResNets resulted in a 3.75% error on the test set. They also performed ResNet analysis on the CIFAR-10 dataset. This data contains 50,000 training images and 10,00- testing images belonging to 10 classes. An average testing error rate of 6.61% was obtained with the 110-layer ResNet He et al. (2015).

Since the ResNet models have been known to work on image recognition tasks He et al. (2015), the ResNet50 (50-layer residual network) was used to learn the features of the ultrasound images before classifying them with the classic machine learning models discussed in 2.2. The pre-trained weights from the ImageNet dataset were used to extract the features.

2.2 Classic Machine-Learning Models

To stay true to Occam's razor (choosing the simpler model over the more complicated model unless the more complicated model is necessary for accuracy improvement Ortner & Leitgeb (2011)), simpler, classic machine learning models were used to classify the two modes (B-Mode and Power Doppler) of US images into the four grades of RA.

2.2.1 *k*-Nearest Neighbors

k-Nearest Neighbors (KNN) is one of the classic machine-learning models used for classifying the features found with ResNet50. This model works by finding the k (a hyperparameter) closest neighbors. The maximum class of those k closest neighbors of a training set is considered the predicted class of the test sample. If there is a tie, the predicted class is chosen randomly. In this model, the Euclidean distance is used to determine the k-nearest neighbors. A visual representation of KNN with $k = 1, k = 2$, and $k = 3$ is shown in Figure 2.1 Tan et al. (2019).

In the left image of ($k = 1$) Figure 2.1, the maximum class is the negative class. Therefore, the test sample is predicted as negative. In the middle image ($k = 2$), there is a tie of classes (i.e., there is one negative neighbor and one positive neighbor), meaning the class is randomly predicted as either positive or negative. For the right image ($k = 3$), there are two positive neighbors and one negative neighbor. Therefore, the test sample is predicted to be positive Tan et al. (2019).

As can be seen in the three examples in 2.1, the value of the hyperparameter k is very important as the predictions change depending on how many neighbors there are. In order to avoid overfitting (learning the noise of the

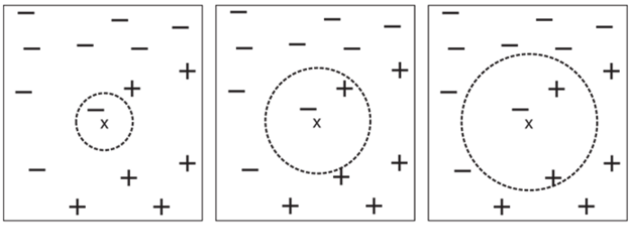


Figure 2.1: Example of KNN with $k = 1$ (left), $k = 2$ (middle), and $k = 3$ (right).

Note. Adapted from Tan et al. (2019)

dataset instead of the underlying distribution) with too small of a k values and to avoid underfitting (not learning enough of the distribution) with too large of a k value, a validation set should be used Tan et al. (2019). By iterating through k values and computing the validation accuracy of each iteration, the optimal k value can be found by determining the k value with the maximum accuracy.

2.2.2 Linear Support Vector Machine

Another classic machine learning model used to classify the features learned in ResNet50 is the Linear Support Vector Machine (SVM). The main goal of linear SVM is to find a hyperplane (a line) that splits the classes. However, linear SVM does not just use any line that will split the classes; it finds the hyperplane that will maximize the margins between the line and the closest data point, called support vectors, on either side. An example is shown in Figure 2.2 Tan et al. (2019).

The basic equation for the hyperplane of the SVM model is $w^T x + b = 0$, as shown in 2.2. x represents the inputs. w represents the weights and is a parameter. b is another parameter and represents the bias Tan et al. (2019).

The margin hyperplanes, the lines parallel to the hyperplane that touch the closest data point of the classes, are equidistant from the hyperplane. In Figure 2.2, the margin hyperplanes are represented as $w^T x + b = -1$ and $w^T x + b = 1$. The distance between the margin hyperplanes and the hyperplane is maximized to allow for greater generalization. Smaller distances would make any slight disturbance in the boundary layer cause a large impact on classification. Therefore, the maximum distance between the hyperplane and margin hyperplanes makes for a more robust and optimal model Tan et al. (2019).

Once the hyperplane that maximizes the distance between the hyperplane and the margin hyperplanes is found, the SVM model can predict classes based

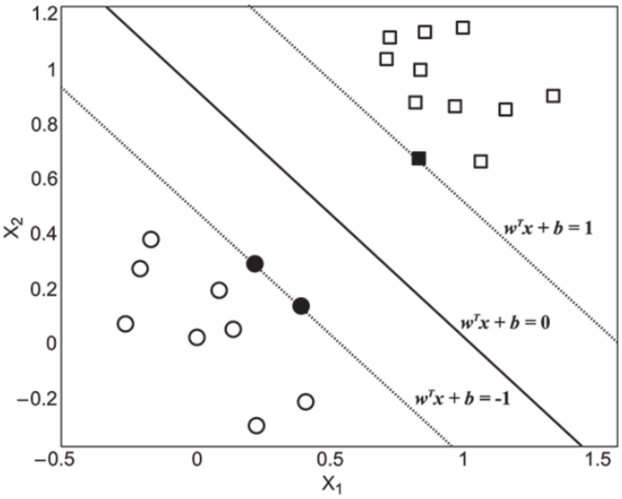


Figure 2.2: Example of a linear SVM model with a linearly separable dataset.

Note. Adapted from Tan et al. (2019)

on the result of $\hat{y}_i = w^T x_i + b$. If $\hat{y}_i$ is negative, x_i belongs to the negative class (shown as circles in Figure 2.2). Reversely, if $\hat{y}_i$ is positive, then x_i is predicted as the positive class (displayed as squares in Figure 2.2) Tan et al. (2019).

Mathematically, the optimization is shown in Equation 2.1, where w and b are parameters, y_i are the true outputs or classes (denoted as -1 and 1), and x_i are the inputs Tan et al. (2019).

$$\begin{aligned} &\min_{w,b} \frac{||w||^2}{2} \\ \text{s. t.} \quad & y_i(w^T x_i + b) \geq 1 \end{aligned} \tag{2.1}$$

Up to this point, there has been an assumption with linear SVM that the classes are linearly separable, meaning one line can perfectly separate the classes. However, this does not happen often in the real world. Therefore, a modification has been made to linear SVM, named soft-margin SVM, that allows for a few samples to be allowed within or on the wrong side of the margins, allowing linear separation on non-linearly separable data and allowing for potentially better generalization. An example of soft-margin SVM is shown in Figure 2.2 Tan et al. (2019).

Soft-margin SVM allows for some samples to be within the margin or be training errors (e.g., a positive sample on the negative side) by introducing a slack variable (ξ) to the constraint. However, by relaxing the constraint, a new cost function must be created to avoid underfitting and large training error. This new cost function introduces a hyperparameter, denoted as C, that is a trade-

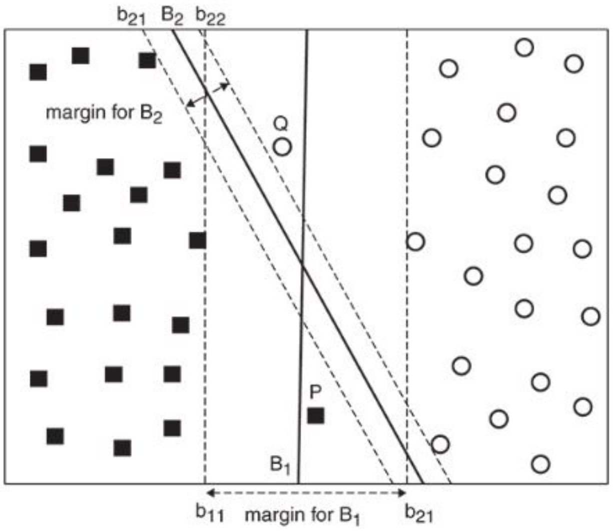


Figure 2.3: Example of a linear soft-margin SVM model.

Note. Adapted from Tan et al. (2019)

off between minimizing the training error (i.e., overfitting) and maximizing the margins (i.e., underfitting). The new optimization problem is shown in Equation 2.2 Tan et al. (2019).

$$
\begin{aligned}
& \min_{w,b,\xi_i} \frac{||w||^2}{2} + C\sum_{i=1}^{n} \xi_i \\
\text{s. t.} \quad & y_i(w^T x_i + b) \geq 1 - \xi_i \\
& \xi_i \geq 0
\end{aligned}
\tag{2.2}
$$

Hyperparameter C was tuned using a validation dataset. The model with the C value that results in the greatest validation accuracy (or smallest validation error) was considered the optimum model.

2.2.3 Non-Linear Support Vector Machine

Soft-margin linear SVM allows for linear separation with some training errors but still assumes a linear separation. Some datasets are not linearly separable but rather separable by a non-linear function, such as a circle. An example of a dataset that requires a nonlinear decision boundary is shown in Figure 2.4 Tan et al. (2019).

This non-linear function can be found with SVM by transforming the true inputs into a different space where it is linearly separable. This transformation is completed using what is called a kernel. Once the linear hyperplane is

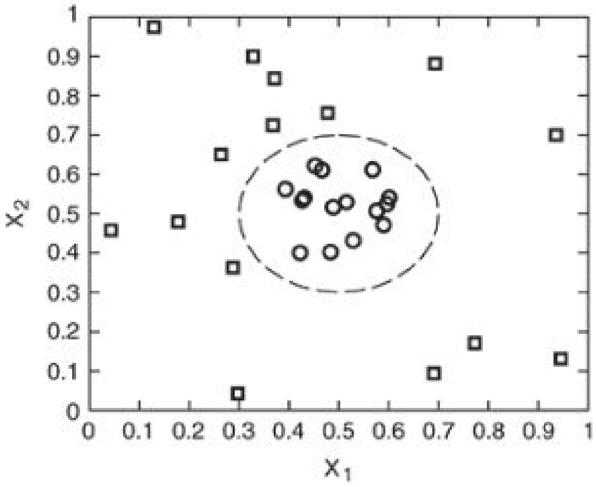


Figure 2.4: Example of a non-linear decision boundary.

Note. Adapted from Tan et al. (2019)

determined in the transformation space, it is translated back to the original input space, resulting in a non-linear decision boundary Tan et al. (2019).

Mathematically, the transformed inputs are written as $\varphi(x_i)$, where the inputs (x_i) are transformed with $\varphi(\cdot)$. In the soft-margin SVM optimization problem (maintaining the ability for some samples to pass through the margins and avoiding overfitting), the x_i's are replaced with $\varphi(x_i)$, as displayed in Equation 2.3 Tan et al. (2019).

$$\begin{aligned} &\min_{w,b,\xi_i} \frac{||w||^2}{2} + C\sum_{i=1}^{n} \xi_i \\ \text{s. t.} \quad & y_i(w^T\varphi(x_i)+b) \geq 1-\xi_i \\ & \xi_i \geq 0 \end{aligned} \tag{2.3}$$

This research uses the radial basis function kernel (RBF) to assist with the nonlinear transformation. This kernel has a hyperparameter, σ (the standard deviation of the kernel). In the code for this research, γ is used instead of σ Tan et al. (2019). Both hyperparameter γ and hyperparameter C were tuned using the validation set's accuracy.

2.2.4 Decision Tree

The next classic machine learning model is the Decision Tree. A decision tree uses hierarchical rules to classify the inputs into a set number of classes. An example of a decision tree is shown in Figure 2.5 Tan et al. (2019).

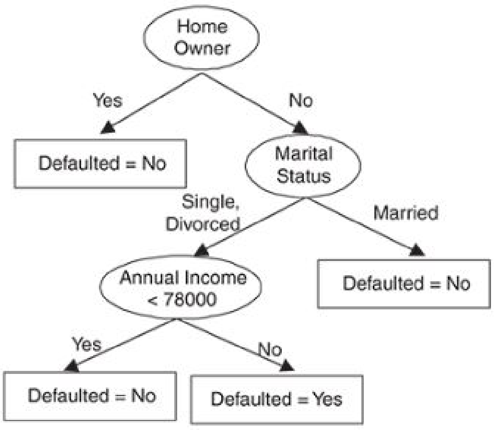


Figure 2.5: Example of a decision tree with a binary classification problem (Defaulted on Loan and Did Not Default on Loan) and the following inputs: Home Ownership (binary), Marital Status (nominal), and Annual Income (continuous).

Note. Adapted from Tan et al. (2019)

The first node (e.g., the circle labeled Home Owner in Figure 2.5) is considered the root node. A leaf or terminal node is a node where there are no outgoing links. They give a class prediction. In Figure 2.5, the leaves are the rectangles labeled “Defaulted = No” or “Defaulted = Yes.” Internal nodes are between the leaf and root node (e.g., the Marital Status circle and Annual Income $<$ 78000 circle in Figure 2.5). The root node and internal nodes have attribute test conditions. The possible outcomes of the attribute test conditions result in a child node. A child node can be a leaf or an internal node Tan et al. (2019).

Hunt’s algorithm is used to decide the structure of the optimum tree. This algorithm is considered greedy because it determines the optimum split at each layer (although the true overall optimum decision tree may not be optimum at each layer). The splitting criterion used for this research was the Gini index (a measure of impurity), shown in Equation 2.4, where c is the number of classes and $p_i(t)$ is the relative frequency of training instances in the current class i and node t Tan et al. (2019). The determination of which attribute to split is chosen by the option that results in the purest node (i.e., a pure node contains samples belonging to only one class).

$$\text{Gini index} = 1 - \sum_{i=0}^{c-1} p_i(t)^2 \tag{2.4}$$

Without hyperparameters, the decision tree will keep splitting until every sample is predicted accurately. Since this causes overfitting issues, the maximum depth (how far down the tree can go) was tuned with the validation set.

2.2.5 Random Forest

The Random Forest model is an ensemble of decision trees. However, each individual tree is not trained on the entire dataset but rather trained on subsets of the data samples and based on a subset of attributes. This allows for the individual trees to be different and to focus on different features Tan et al. (2019).

A test sample goes through all the trees. Each tree gives a prediction for the sample. The maximum predicted class of the trees is the predicted class Tan et al. (2019).

The Gini index was used once again to calculate the impurity. The hyperparameters tuned for this model include the number of trees, the maximum depth of the trees, and the maximum features at each split.

2.2.6 Neural Network

The distinction between this particular neural network and the neural networks described in Chapter 3 is that this neural network, known as the Multi-Layer Perceptron (MLP), is in the Python sci-kit learn library. Whereas, the neural networks used in Chapter 3 are developed in the PyTorch library.

An example of a basic perceptron architecture is shown in Figure 2.6. In this example, there are three inputs and one output. Each input, x_i, is multiplied by a weight, w_i, represented by an arrow connecting the inputs to the neuron. The weighted inputs are added together with the addition of a bias, b. This weighted sum is put through a non-linear function, called the activation layer. Figure 2.6 denotes a sign function (-1 if the given value is negative and 1 if the given value is positive). The result of the activation layer is the output of the neuron, y Tan et al. (2019).

Figure 2.7 shows an MLP with p inputs and L layers. Each layer has an arbitrary number of neurons. The output of the neurons of each layer are a weighted sum of the previous hidden layer's neurons (or inputs) plus a bias term, which is put through an activation layer, as discussed in the basic architecture of a perceptron Tan et al. (2019). The ReLu function was used for the activation layers. Figure 2.7 shows an example with a single final output.

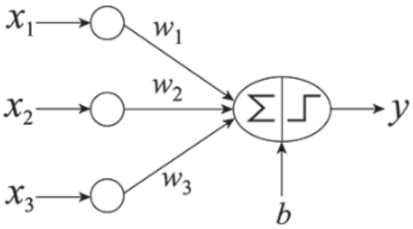


Figure 2.6: Example of a perceptron architecture.

Note. Adapted from Tan et al. (2019)

However, four outputs were used for this research because there are four grades of RA. The validation set was used to tune the size of the hidden layers.

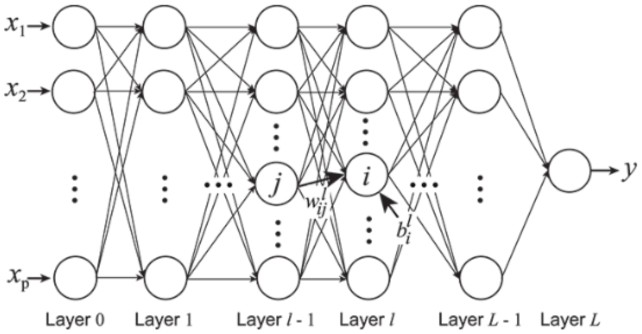


Figure 2.7: Multi-Layer Perceptron with one output.

Note. Adapted from Tan et al. (2019)

2.2.7 Adaboost

Adaboost is an ensemble method. This algorithm trains k classifiers and weights them based on hard-to-classify samples. In other words, instead of predicting the class with the maximum prediction among the classifiers, a weight is placed on each classifier to determine its importance in the final decision Tan et al. (2019).

Equation 2.5 shows the equation for updating the weights, where $\alpha_i = \frac{1}{2}ln(\frac{1-\varepsilon_i}{\varepsilon_i})$ and Z_j ensures the sum of weights equals one Tan et al. (2019). Since α_i requires the error, ε_i, for calculation, the weights are dependent on the error rate. If the error rate is high for a certain classifier, it is punished with a smaller weight. Therefore, Adaboost is useful in predicting hard-to-classify samples Tan et al. (2019).

$$w_i^{(j+1)} = \frac{w_i^{(j)}}{Z_j} * \begin{cases} e^{-\alpha_j} & \text{if } C_j(x_i) = y_i \\ e^{\alpha_j} & \text{if } C_j(x_i) = y_i \end{cases} \tag{2.5}$$

Equation 2.6 is the formula for calculating the error. $I(p)$ is one when the class is predicted correctly and zero when the class is not predicted correctly. Since the weight update function relies on ε_i and the error function relies on w_j, initial weights are required. Adaboost assumes equal weight for all classifiers at first, which is $\frac{1}{k}$ for k classifiers Tan et al. (2019).

$$\varepsilon_i = \frac{1}{N}[\sum_{j=1}^{N} w_j I(C_i(x_j) \neq y_j)] \tag{2.6}$$

The decision tree classifier was used for all classifiers, and the hyperparameter that was tuned was the number of estimators used before boosting was terminated (unless there was a perfect fit).

2.2.8 Quadratic Discriminant Analysis

The quadratic discriminant analysis (QDA) is the eighth classic machine learning model used to predict sonographic RA synovitis grades. QDA works similarly to linear discriminant analysis by computing discriminant scores, except QDA does not assume identical covariance matrices for all classes. The formula for the discriminant scores is shown in Equation 2.7, where Σ_k is the k class covariance matrix, μ_k is the mean vector for the k class, and π_k is the prior probability QDA.

$$\delta_k(x) = -\frac{1}{2} log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) + log(\pi_k) \tag{2.7}$$

For each sample, a discriminant score is calculated for each class. The class with the largest discriminant score is considered the predicted class. The hyperparameter tuned here was the regularization parameter.

2.3 Results

Table 2.1 shows the B-Mode results of the best model found in each of the eight classic machine-learning models by tuning the hyperparameters on the validation dataset. These results were unsatisfactory as the average testing accuracy of the eight models was only 38.31%. If the model only predicted the largest class, Grade 2 (the simplest model), there would be an approximately

	Validation Accuracy	Testing Accuracy
KNN	39.95%	39.03%
Linear SVM	42.63%	41.29%
Non-Linear SVM	44.24%	40.32%
Decision Tree	36.19%	35.48%
Random Forest	43.70%	37.74%
Neural Network	38.87%	33.55%
Adaboost	36.73%	36.45%
QDA	40.48%	35.16%

Table 2.1: Validation and testing accuracies for the eight classic machine learning models on the B-Mode US images.

37.47% accuracy. Satisfactory results would have a larger improvement in accuracy from the simplest model than seen here.

Table 2.2 shows the Power Doppler results of the best model found in each of the eight classic machine-learning models by tuning the hyperparameters on the validation dataset. These results were unsatisfactory as the average testing accuracy of the eight models was only 50.82%. If the model only predicted the largest class, Grade 0 (the simplest model), there would be an approximately 50.37% accuracy. Satisfactory results would have a larger improvement in accuracy from the simplest model than seen here.

	Validation Accuracy	Testing Accuracy
KNN	51.51%	50.66%
Linear SVM	48.49%	52.96%
Non-Linear SVM	49.59%	52.30%
Decision Tree	50.14%	50.66%
Random Forest	54.52%	50.33%
Neural Network	50.41%	47.04%
Adaboost	51.51%	49.34%
QDA	45.75%	50.33%

Table 2.2: Validation and testing accuracies for the eight classic machine learning models on the Power Doppler US images.

Chapter 3

Transfer Learning with Deep Neural Networks

Since the classic machine learning models did not provide satisfactory results, deep learning models were developed to determine the added model complexity's effect on the sonographic RA synovitis grading classification ability.

3.1 Vanilla Fine-Tuning: ResNet50v2

Before fine-tuning ResNet50, the base model was switched to ResNet50v2. The ResNet50v2 model was developed by He et al. (2016). He et al. (2016) sought to increase generalization and ease the training of deep residual networks by "... using identity mappings as the skip connections and after-addition activation." The use of identity mapping allowed for direct propagation, which eases the optimization. The after-addition activation allowed for more generalization by normalizing the input before going into the next residual block, easing the training process more. Furthermore, ResNet50v2 was trained on a larger ImageNet dataset (ImageNet21K), which contains 20.000 more classes than the ImageNet dataset ResNet50 trained on Wightman (2019). The number of epochs (up to 100) was tuned using the validation set.

3.2 Results of Vanilla Fine-Tuning

For the deep learning models, the top 1, top 2, and top 3 class success rates were used to evaluate the model overall. The top 1 success rate is equivalent to the classification accuracy used to evaluate the classic machine learning models. Since there are no crystal clear diagnostic criteria for RA synovitis grades, the top 2 success rates were evaluated to account for the borderline cases between two RA synovitis grades. The top 3 success rates were evaluated to confirm the model's predicting ability. The top 3 success rates should be above 85% to verify the model's ability to classify. The quadratic weighted kappa statistics were used as a secondary comparative statistic sci (2023). Since stochastic gradient descent was used for the optimization of the ResNet50v2 models, ten trial runs were conducted to account for the inherent randomness associated with stochastic gradient descent, meaning the runs will not be exactly the same even with the use of a manual seed. The mean and standard deviation of the ten trial runs are given for all deep learning model results.

3.2.1 B-Mode

The top 1 validation mean success rate for the Vanilla Fine-Tuning model with B-Mode US images was 42.23%. This was a 1.88% increase from the average of the eight classic machine learning models' validation classification accuracy. The mean top 2 validation success rate was 70.78%, while the mean top 3 success rate was 89.87%. There was a 28.55% mean increase from the top 1 and top 2 success rates, meaning there were several borderline cases. The mean top 3 success rate was above the 85% cutoff for a decent model (Table 3.1).

	Validation Success Rate
Top 1	42.23% +/- 1.98%
Top 2	70.78% +/- 2.74%
Top 3	89.87% +/- 1.43%

Table 3.1: Mean and standard deviation of the validation top 1, top 2, and top 3 success rates for ten trial runs of the Vanilla Fine-Tuning model with B-Mode US images.

The testing mean top 1 success rate was 41.03%, which was a 2.72% increase from the average testing classification accuracy of the eight classic machine learning models. The testing mean top 2 success rate was 72.58%, while the top 3 was 91.06%. There was a 31.55% increase from top 1 to top 2, meaning there were several borderline cases. Additionally, the top 3 testing success rate was above the 85% cutoff (Table 3.2). There was an improvement in accuracy from the classic machine learning models, but a larger improvement from the simplest model would be more satisfactory.

The mean quadratic weighted kappa statistic for the validation set was 0.2436, whereas the mean weighted kappa statistic for the testing set was

	Testing Success Rate
Top 1	41.03% +/- 1.62%
Top 2	72.58% +/- 2.24%
Top 3	91.06% +/- 1.32%

Table 3.2: Mean and standard deviation of the testing top 1, top 2, and top 3 success rates for ten trial runs of the Vanilla Fine-Tuning model with B-Mode US images.

0.2564. Ideally, the kappa statistic would be closer to 1 (Table 3.3). Therefore, these results were unsatisfactory.

	Quadratic Weighted Kappa Statistic
Validation	0.2436 +/- 0.0247
Testing	0.2564 +/- 0.0400

Table 3.3: Mean and standard deviation of the quadratic weighted kappa statistic for the validation and testing sets for ten trial runs of the Vanilla Fine-Tuning model with B-Mode US images.

Tables 3.4, 3.5, and 3.6 show the mean and standard deviation of the top 1, top 2, and top 3 successes, failures, and sensitivity by sonographic RA synovitis grades for the validation set, respectively. Overall, grades 0 and 2 had the highest sensitivity and largest number of successes than the other grades, potentially due to the fact that grades 0 and 2 were the largest classes and are the more commonly diagnosed RA synovitis grades. Furthermore, grade 1 had the lowest sensitivity for top 1, top 2, and top 3, meaning it was the hardest class to classify.

	Successes	Failures	Sensitivity
Grade 0	57.80 +/- 9.63	58.20 +/- 9.63	49.83% +/- 8.30%
Grade 1	3.60 +/- 1.80	38.40 +/- 1.80	8.57% +/- 4.29%
Grade 2	82.4 +/- 18.06	64.60 +/- 18.06	56.05% +/- 12.29%
Grade 3	13.70 +/- 3.29	54.30 +/- 3.29	20.15% +/- 4.84%

Table 3.4: The validation set's Top 1 Successes, Failures, and Sensitivity per grade for the B-Mode Vanilla Fine-Tuning model.

Tables 3.7, 3.8, and 3.9 show the mean and standard deviation of the top 1, top 2, and top 3 successes, failures, and sensitivity by sonographic RA synovitis grades for the testing set, respectively. Once again, grades 0 and

	Successes	Failures	Sensitivity
Grade 0	91.50 +/- 7.55	24.50 +/- 7.55	78.88% +/- 6.51%
Grade 1	13.90 +/- 4.99	28.10 +/- 4.99	33.10% +/- 11.88%
Grade 2	123.80 +/- 12.94	23.20 +/- 12.94	84.22% +/- 8.81%
Grade 3	34.80 +/- 5.23	33.20 +/- 5.23	51.18% +/- 7.69%

Table 3.5: The validation set's Top 2 Successes, Failures, and Sensitivity per grade for the B-Mode Vanilla Fine-Tuning model.

	Successes	Failures	Sensitivity
Grade 0	110.20 +/- 3.37	5.80 +/- 3.37	95.00% +/- 2.91%
Grade 1	31.60 +/- 2.97	10.40 +/- 2.97	75.24%+/- 7.08%
Grade 2	141.10 +/- 4.66	5.90 +/- 4.66	95.99% +/- 3.17%
Grade 3	52.30 +/- 2.93	15.70 +/- 2.93	76.91% +/- 4.32%

Table 3.6: The validation set's Top 3 Successes, Failures, and Sensitivity per grade for the B-Mode Vanilla Fine-Tuning model.

2 had the highest sensitivity and largest number of successes than the other grades. Furthermore, grade 1 had the lowest sensitivity for top 1, 2, and 3, meaning it was the hardest class to classify for the testing set too.

	Successes	Failures	Sensitivity
Grade 0	49.90 +/- 10.47	54.10 +/- 10.47	47.98% +/- 10.07%
Grade 1	2.30 +/- 1.79	36.70 +/- 1.79	5.90% +/- 4.59%
Grade 2	63.10 +/- 12.64	52.90 +/- 12.64	54.40% +/- 10.89%
Grade 3	11.90 +/- 4.72	39.10 +/- 4.72	23.33% +/- 9.26%

Table 3.7: The test set's Top 1 Successes, Failures, and Sensitivity per grade for the B-Mode Vanilla Fine-Tuning model.

3.2.2 Power Doppler

The mean top 1 validation success rate was 63.26%, which was a 13.02% increase from the average classic machine learning model's validation accuracy. The mean top 2 validation success rate was 81.15%, while the mean top 3 validation success rate was 94.90% (Table 3.10). There was a mean increase of 17.89% from top 1 to top 2 validation success rate, showing there were several borderline cases. Furthermore, the mean top 3 validation success rate was well above the 85% threshold, further confirming the validity of the model.

	Successes	Failures	Sensitivity
Grade 0	83.90 +/- 8.28	20.10 +/- 8.28	80.67% +/- 7.96%
Grade 1	13.70 +/- 5.35	25.30 +/- 5.35	35.13% +/- 13.71%
Grade 2	99.20 +/- 8.63	16.80 +/- 8.63	85.52% +/- 7.44%
Grade 3	28.20 +/- 4.53	22.80 +/- 4.53	55.29% +/- 8.89%

Table 3.8: The test set's Top 2 Successes, Failures, and Sensitivity per grade for the B-Mode Vanilla Fine-Tuning model.

	Successes	Failures	Sensitivity
Grade 0	98.90 +/- 2.77	5.10 +/- 2.77	95.10% +/- 2.67%
Grade 1	30.80 +/- 3.03	8.20 +/- 3.03	78.97% +/- 7.76%
Grade 2	112.50 +/- 2.66	3.5 +/- 2.66	96.98% +/- 2.29%
Grade 3	40.10 +/- 2.62	10.90 +/- 2.62	78.63% +/- 5.15%

Table 3.9: The test set's Top 3 Successes, Failures, and Sensitivity per grade for the B-Mode Vanilla Fine-Tuning model.

	Validation Success Rate
Top 1	63.26% +/- 1.15%
Top 2	81.15% +/- 1.49%
Top 3	94.90% +/- 1.12%

Table 3.10: Mean and standard deviation of the validation top 1, top 2, and top 3 success rates for ten trial runs of the Vanilla Fine-Tuning model with Power Doppler US images.

The mean top 1 testing success rate was 61.78%, which was a 10.96% improvement from the average testing classification accuracy of the eight classic machine learning models. The mean top 2 testing success rate was 80.79%, and the mean top 3 testing success rate was 94.21%, as shown in Table 3.11. There was a 19.01% mean increase from the top 1 to the top 2 testing success rate, meaning there were several borderline cases. Additionally, the top 3 testing success rate further validated the model by exceeding the 85% threshold. The top 1 testing and validation success rates were satisfactory, meaning the Vanilla Fine-Tuning on the Power Doppler US images model was an acceptable model for predicting sonographic RA synovitis grades.

	Testing Success Rate
Top 1	61.78% +/- 1.69%
Top 2	80.79% +/- 1.81%
Top 3	94.21% +/- 1.31%

Table 3.11: Mean and standard deviation of the testing top 1, top 2, and top 3 success rates for ten trial runs of the Vanilla Fine-Tuning model with Power Doppler US images.

Since the quadratic weighted kappa statistic was closer to 1 than 0 for both the validation and testing datasets, the Vanilla Fine-Tuning model on the Power Doppler US images was validated as an acceptable model.

	Quadratic Weighted Kappa Statistic
Validation	0.6701 +/- 0.2320
Testing	0.6106 +/- 0.0386

Table 3.12: The mean and standard deviation of the quadratic weighted kappa statistic for the validation and testing sets with ten trial runs of the Vanilla Fine-Tuning model with Power Doppler US images.

Tables 3.13, 3.14, and 3.15 show the mean and standard deviation for the top 1, top 2, and top 3 breakdowns of successes, failures, and sensitivity by RA synovitis grades on the validation dataset, respectively. Overall, grade 0 had the highest sensitivity, potentially due to grade 0 being the largest class. Grades 2 and 3 had the next highest sensitivity. Lastly, grade 1 had the lowest sensitivity, making it the hardest grade to classify.

Tables 3.16, 3.17, and 3.18 show the mean and standard deviation of the top 1, top 2, and top 3 breakdowns of successes, failures, and sensitivity by RA synovitis grades on the testing dataset, respectively. Once again, grade 0 had

	Successes	Failures	Sensitivity
Grade 0	147.10 +/- 5.26	19.90 +/- 5.26	88.08% +/- 3.15%
Grade 1	1.80 +/- 2.14	43.20 +/- 2.14	4.00% +/- 4.75%
Grade 2	35.90 +/- 6.25	49.10 +/- 6.25	42.24% +/- 7.36%
Grade 3	46.10 +/- 4.13	21.90 +/- 4.13	67.79% +/- 6.08%

Table 3.13: The validation set's Top 1 Successes, Failures, and Sensitivity per grade for the Power Doppler Vanilla Fine-Tuning model.

	Successes	Failures	Sensitivity
Grade 0	156.90 +/- 3.56	10.10 +/- 3.56	93.95% +/- 2.13%
Grade 1	14.70 +/- 8.40	30.30 +/- 8.40	32.67% +/- 18.67%
Grade 2	68.10 +/- 10.07	16.90 +/- 10.07	80.12% +/- 11.85%
Grade 3	56.50 +/- 3.07	11.50 +/- 3.07	83.09% +/- 4.52%

Table 3.14: The validation set's Top 2 Successes, Failures, and Sensitivity per grade for the Power Doppler Vanilla Fine-Tuning model.

	Successes	Failures	Sensitivity
Grade 0	162.20 +/- 2.71	4.80 +/- 2.71	97.13% +/- 1.62%
Grade 1	39.50 +/- 7.14	5.50 +/- 7.14	87.78% +/- 15.88%
Grade 2	82.50 +/- 1.69	2.50 +/- 1.69	97.06% +/- 1.99%
Grade 3	62.20 +/- 1.99	5.80 +/- 1.99	91.47% +/- 2.93%

Table 3.15: The validation set's Top 3 Successes, Failures, and Sensitivity per grade for the Power Doppler Vanilla Fine-Tuning model.

the highest sensitivity, while grades 2 and 3 had the next highest sensitivity. Furthermore, grade 1 had the lowest sensitivity.

	Successes	Failures	Sensitivity
Grade 0	131.90 +/- 6.70	21.10 +/- 6.70	86.21% +/- 4.38%
Grade 1	1.80 +/- 1.54	35.20 +/- 1.54	4.86% +/- 4.15%
Grade 2	23.20 +/- 4.98	41.80 +/- 4.98	35.69% +/- 7.66%
Grade 3	30.90 +/- 4.16	18.10 +/- 4.16	63.06% +/- 8.49%

Table 3.16: The test set's Top 1 Successes, Failures, and Sensitivity per grade for the Power Doppler Vanilla Fine-Tuning model.

	Successes	Failures	Sensitivity
Grade 0	140.10 +/- 4.89	12.90 +/- 4.89	91.57% +/- 3.19%
Grade 1	15.30 +/- 7.93	21.70 +/- 7.93	41.35% +/- 21.42%
Grade 2	49.20 +/- 8.63	15.80+/- 8.63	75.69% +/- 13.28%
Grade 3	41.00 +/- 1.67	8.00 +/- 1.67	83.67% +/- 3.41%

Table 3.17: The test set's Top 2 Successes, Failures, and Sensitivity per grade for the Power Doppler Vanilla Fine-Tuning model.

	Successes	Failures	Sensitivity
Grade 0	145.80 +/- 4.49	7.20 +/- 4.49	95.29% +/- 2.93%
Grade 1	32.40 +/- 5.33	4.60 +/- 5.33	87.57% +/- 14.41%
Grade 2	62.80 +/- 2.68	2.20 /- 2.68	96.62% +/- 4.12%
Grade 3	45.40 +/- 1.28	3.60 +/- 1.28	92.65% +/- 2.61%

Table 3.18: The test set's Top 3 Successes, Failures, and Sensitivity per grade for the Power Doppler Vanilla Fine-Tuning model.

3.3 Domain-Adaptive Tuning

Although the Vanilla Fine-Tuning model on the Power Doppler US images proved to be an acceptable model, further improvement was plausible. Therefore, a Domain-Adaptive Tuning model was proposed.

3.3.1 Theory

Since Power Doppler US images are similar to B-Mode but with an added color channel Babcock et al. (1996), a Domain-Adaptive Tuning model was proposed. This model works by fine-tuning the ResNet50v2 model on the entire B-Mode dataset (combining the training, validation, and testing sets). The resulting fine-tuned ResNet50v2 model was considered the pre-trained model for the Power Doppler US images. The training, validation, and testing are performed the same as the Vanilla Fine-Tuning model for the Power Doppler US images but with the B-Mode fine-tuned ResNet50v2 model instead of the original ResNet50v2 model pre-trained on ImageNet21K. The theory is that the ResNet50v2 model would learn the features behind the color channel first, and then learn the specific features seen in the Power Doppler US images, potentially improving the accuracy of predicting the sonographic RA synovitis grades of Power Doppler US images. Once again, the number of epochs (up to 100) was tuned using the validation set.

3.3.2 Results

As seen in Table 3.19, the mean top 1 validation success rate was 61.95%, which was a 1.31% decrease from the Vanilla Fine-Tuning model's mean top 1 validation success rate on Power Doppler US images. The mean top 2 validation success rate was 80.41% (a 0.74% decrease from the Vanilla Fine-Tuning model), while the mean top 3 validation success rate was 94.44% (a 0.46% decrease from the Vanilla Fine-Tuning model). Since the mean top 1, top 2, and top 3 validation success rates were not as high as the success rates of the Vanilla Fine-Tuning model on the Power Doppler US images, the Domain-Adaptive Tuning model was not considered a more satisfactory model than the Vanilla Fine-Tuning model.

	Validation Success Rate
Top 1	61.95% +/- 1.14%
Top 2	80.41% +/- 1.00%
Top 3	94.44% +/- 1.06%

Table 3.19: Mean and standard deviation of the validation top 1, top 2, and top 3 success rates for ten trial runs of the Domain-Adaptive Tuning model.

In Table 3.20, the mean top 1 testing success rate was calculated to be 61.68%, which was 0.10% less than the mean top 1 testing success rate of the Power Doppler Vanilla Fine-Tuning model. The mean top 2 testing success rate was 79.97%, which was a 0.82% decrease from the Vanilla Fine-Tuning model's mean top 2 testing success rate. Additionally, the mean top 3 testing success rate was 93.39% (a 0.82% decrease from the Vanilla Fine-Tuning model's mean top 3 testing success rate). Once again, the slight decrease in top 1, top 2, and top 3 testing success rates showed that the Domain-Adaptive Tuning model did not significantly improve the predicting ability of RA synovitis grading.

	Testing Success Rate
Top 1	61.68% +/- 1.57%
Top 2	79.97% +/- 0.98%
Top 3	93.39% +/- 1.41%

Table 3.20: Mean and standard deviation of the testing top 1, top 2, and top 3 success rates for ten trial runs of the Domain-Adaptive Tuning model.

The mean quadratic weighted kappa statistic for the validation set was 0.6721, while the mean quadratic weighted kappa statistic for the testing set was 0.6113. These were both slightly higher than the respective kappa statistics of the Power Doppler Vanilla Fine-Tuning model. Since there was little improvement from the vanilla model, the Domain-Adaptive Tuning model was not considered a more satisfactory model.

	Quadratic Weighted Kappa Statistic
Validation	0.6721 +/- 0.0229
Testing	0.6113 +/- 0.0190

Table 3.21: Mean and standard deviation of the quadratic weighted kappa statistic for the validation and testing sets for ten trial runs of the Domain-Adaptive Tuning model.

Tables 3.22, 3.23, and 3.24 display the mean and standard deviation of the top 1, top 2, and top 3 validation successes, failures, and sensitivity by RA synovitis grade, respectively. In all three cases, grade 0 had the highest sensitivity followed by grade 3 and grade 2. Once again, grade 1 had the lowest sensitivity and proved to be the most difficult class to predict accurately.

Tables 3.25, 3.26, and 3.27 display the mean and standard deviation of the top 1, top 2, and top 3 testing successes, failures, and sensitivity by RA synovitis grade, respectively. In all three cases, grade 0 had the highest sensitivity followed by grade 3 and grade 2. Furthermore, grade 1 had the lowest sensitivity.

	Successes	Failures	Sensitivity
Grade 0	146.10 +/- 4.83	20.90 +/- 4.83	87.49% +/- 2.89%
Grade 1	1.10 +/- 0.94	43.90 +/- 0.94	2.44% +/- 2.10%
Grade 2	27.20 +/- 6.14	57.80 +/- 6.14	32.00% +/- 7.23%
Grade 3	51.70 +/- 2.87	16.30 +/- 2.87	76.03% +/- 4.21%

Table 3.22: The validation set's Top 1 Successes, Failures, and Sensitivity per grade for the Power Doppler Vanilla Fine-Tuning model.

	Successes	Failures	Sensitivity
Grade 0	156.90 +/- 2.26	10.10 +/- 2.26	93.95% +/- 1.35%
Grade 1	12.20 +/- 1.60	32.80 +/- 1.60	27.11% +/- 3.56%
Grade 2	66.50 +/- 5.63	18.50 +/- 5.63	78.24% +/- 6.62%
Grade 3	57.90 +/- 2.12	10.10 +/- 2.12	85.15% +/- 3.12%

Table 3.23: The validation set's Top 2 Successes, Failures, and Sensitivity per grade for the Power Doppler Vanilla Fine-Tuning model.

	Successes	Failures	Sensitivity
Grade 0	163.20 +/- 1.40	3.80 +/- 1.40	97.72% +/- 0.84%
Grade 1	38.00 +/- 2.86	7.00 +/- 2.86	84.44% +/- 6.36%
Grade 2	79.70 +/- 2.28	5.30 +/- 2.28	93.76% +/- 2.69%
Grade 3	63.80 +/- 1.25	4.20 +/- 1.25	93.82% +/- 1.84%

Table 3.24: The validation set's Top 3 Successes, Failures, and Sensitivity per grade for the Power Doppler Vanilla Fine-Tuning model.

	Successes	Failures	Sensitivity
Grade 0	131.60 +/- 5.02	21.40 +/- 5.02	86.01% +/- 3.28%
Grade 1	4.20 +/- 1.17	32.80 +/- 1.17	11.35% +/- 3.15%
Grade 2	20.30 +/- 4.29	44.70 +/- 4.29	31.23% +/- 6.60%
Grade 3	31.40 +/- 2.29	17.60 +/- 2.29	64.08% +/- 4.67%

Table 3.25: The test set's Top 1 Successes, Failures, and Sensitivity per grade for the Power Doppler Vanilla Fine-Tuning model.

	Successes	Failures	Sensitivity
Grade 0	141.00 +/- 2.53	12.00 +/- 2.53	92.16% +/- 1.65%
Grade 1	15.60 +/- 2.15	21.40 +/- 2.15	42.16% +/- 5.82%
Grade 2	45.50 +/- 3.07	19.50 +/- 3.07	70.00% +/- 4.73%
Grade 3	41.00 +/- 1.95	8.00 +/- 1.95	83.67% +/- 3.98%

Table 3.26: The test set's Top 2 Successes, Failures, and Sensitivity per grade for the Power Doppler Vanilla Fine-Tuning model.

	Successes	Failures	Sensitivity
Grade 0	147.40 +/- 2.80	5.60 +/- 2.80	96.34% +/- 1.83%
Grade 1	29.90 +/- 2.47	7.10 +/- 2.47	80.81% +/- 6.67%
Grade 2	60.30 +/- 2.28	4.70 +/- 2.28	92.77% +/- 3.51%
Grade 3	46.30 +/- 1.55	2.70 +/- 1.55	94.49% +/- 3.17%

Table 3.27: The test set's Top 3 Successes, Failures, and Sensitivity per grade for the Power Doppler Vanilla Fine-Tuning model.

3.4 Multitask Transfer Learning

The Vanilla Fine-Tuning on Power Doppler US images and the Domain-Adaptive Tuning models provided satisfactory results. However, the Vanilla Fine-Tuning model on B-Mode US images did not provide satisfactory results and remained a cause of concern. To resolve this concern, a Multitask Transfer Learning (MTL) model was developed.

3.4.1 Theory and Model

An MTL model learns multiple related tasks simultaneously, allowing multiple data sources to be combined. By allowing multiple data source inputs related to the multiple learned tasks or objectives, the MTL model solves some data scarcity issues by combining datasets. Utilizing the multiple datasets, the MTL model shares some or all knowledge between the desired learning tasks Zhang & Yang (2021).

In this research, the B-Mode US training images and the Power Doppler US images are considered the multiple data sources, and their respective sonographic RA synovitis grades are considered the multiple learning tasks or objectives.

The model developed here sought to share some knowledge between the two different modes by using what will be referred to as the Common Net. The

Common Net is the ResNet50v2 model with an output size of 10. Both B-Mode and Power Doppler US images are utilized to learn the weights, allowing the model to learn common features between the two modes. The model then splits into two symmetrical neural networks, B-Net and D-Net. Each of the individual networks comprises of three layers: an input layer of size 10 to match the output layer of the Common Net, a hidden layer of size 100, and an output layer of size 4 to match the number of sonographic RA synovitis grades. The B-Net learns the specific features of the B-Mode US images, while the D-Net learns the specific features of the Power Doppler US images. A visual representation of this model is shown in Figure 3.1.

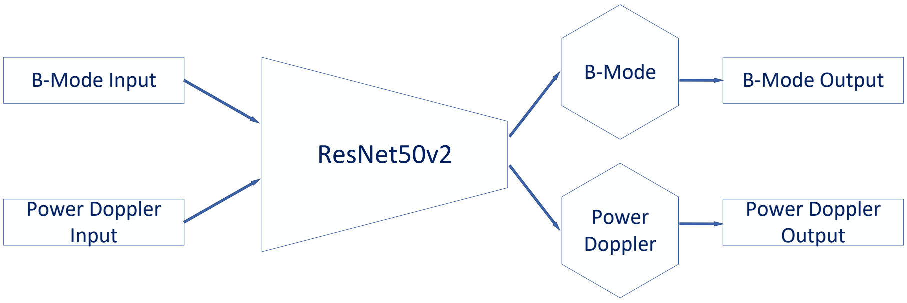


Figure 3.1: MTL model used to simultaneously classify B-Mode and Power Doppler US images into their respective sonographic RA synovitis grades.

The loss function for the MTL model was calculated as shown in Equation 3.1, where $Loss_B$ is the Cross-Entropy Loss of the B-Mode US images and $Loss_D$ is the Cross-Entropy Loss of the Power Doppler US images. The validation set was used to tune the number of epochs (up to 100).

$$Loss = 0.5 * Loss_B + 0.5 * Loss_D \tag{3.1}$$

3.4.2 Results

The mean and standard deviation of the validation success rates for ten trial runs with B-Mode US images are shown in Table 3.28. The top 1 success rate increased from a mean of 42.23% with the Vanilla Fine-Tuning model to a mean of 47.01%. This was a mean increase of 4.78%. The MTL top 2 success rate had a mean increase of 4.39% from the Vanilla Fine-Tuning model, whereas the top 3 success rate had a mean increase of 2.97%. Therefore, the MTL model had an overall positive effect on the predicting ability of sonographic RA synovitis grading on B-Mode images with regards to the top 1, top 2, and top 3 validation success rates.

	Validation Success Rate
Top 1	47.02% +/- 1.75%
Top 2	75.17% +/- 3.10%
Top 3	92.84% +/- 0.98%

Table 3.28: Mean and standard deviation of the validation top 1, top 2, and top 3 success rates for ten trial runs using the B-Mode US images on the MTL model.

The mean and standard deviation of the testing success rates for ten trial runs with B-Mode US images are displayed in Table 3.29. The MTL model's mean top 1 testing success rate was 47.68%, which was a 6.65% mean increase from the Vanilla Fine-Tuning model. The mean top 2 testing success rate increased by 5.10% from the Vanilla Fine-Tuning model, while the mean top 3 success rate increased by 2.68%. Once again, the MTL showed a positive effect on the predicting ability of sonographic RA synovitis grading on B-Mode images with regards to the top 1, top 2, and top 3 mean testing success rates.

	Testing Success Rate
Top 1	47.68% +/- 1.92%
Top 2	77.68% +/- 2.12%
Top 3	93.74% +/- 1.97%

Table 3.29: Mean and standard deviation of the testing top 1, top 2, and top 3 success rates for ten trial runs using the B-Mode US images on the MTL model.

While the Power Doppler US images showed satisfactory results in the Vanilla Fine-Tuning model and the Domain-Adaptive Tuning model, it was imperative to review the results of the MTL model on the Power Doppler US images too. The mean of the top 1 validation success rate with the MTL model (Table 3.30) was slightly less than the mean of the top 1 validation success rates of the Vanilla Fine-Tuning model (Table 3.10), whereas the means of the top 2 and top 3 validation success rates (Table 3.30) were slightly higher than the Vanilla Fine-Tuning model (Table 3.10). The mean of the top 1 validation success rate (Table 3.30) was slightly lower than the mean of the top 1 validation success rate of the Domain-Adaptive Tuning model (Table 3.19), whereas the means of the top 2 and top 3 validation success rates (Table 3.30) were slightly higher than the Domain-Adaptive Tuning model's top 2 and top 3 mean validation success rates (Table 3.19). Therefore, there was no added benefit in regard to the predicting ability of the MTL model in comparison to the Vanilla Fine-Tuning and Domain-Adaptive Tuning models with respect to the mean validation success rates.

	Validation Success Rate
Top 1	61.23% +/- 3.03%
Top 2	82.03% +/- 1.77%
Top 3	95.21% +/- 1.17%

Table 3.30: Mean and standard deviation of the validation top 1, top 2, and top 3 success rates for ten trial runs using the Power Doppler US images on the MTL model.

The testing top 1 success rate mean (Table 3.31) was slightly lower than the Vanilla Fine-Tuning model's top 1 testing mean success rate (Table 3.11), whereas the testing top 2 and top 3 success rates (Table 3.31) than the Vanilla Fine-Tuning model (Table 3.11). The mean of the top 1, top 2, and top 3 testing success rates (Table 3.31) were slightly higher than the Domain-Adaptive Tuning model's top 1, top 2, and top 3 mean testing success rates. Once again, there was no significant benefit to the predicting ability of the MTL model in comparison to the Vanilla Fine-Tuning and Domain-Adaptive Tuning models with regard to the testing success rates of Power Doppler US images.

	Testing Success Rate
Top 1	61.74% +/- 3.11%
Top 2	81.81% +/- 2.27%
Top 3	95.13% +/- 1.66%

Table 3.31: Mean and standard deviation of the testing top 1, top 2, and top 3 success rates for ten trial runs using the Power Doppler US images on the MTL model.

The mean quadratic weighted kappa statistic for the B-Mode validation dataset was 0.3556, which was 0.1120 higher than the Vanilla Fine-Tuning model on B-Mode US images. Additionally, the mean quadratic weighted kappa statistic for the B-Mode testing set was 0.4304, a 0.1740 increase from the Vanilla Fine-Tuning model. These kappa statistics further support the conclusion that the MTL model is a stronger model than the Vanilla Fine-Tuning model for B-Mode US images.

On the other hand, the mean validation quadratic weighted kappa statistic for the Power Doppler US image was 0.7136, as shown in Table 3.33, only

	Quadratic Weighted Kappa Statistic
Validation	0.3556 +/- 0.0183
Testing	0.4304 +/- 0.0190

Table 3.32: The quadratic weighted kappa statistic for the validation and testing sets using the B-Mode US images on the MTL model.

a 0.0435 increase from the Vanilla Fine-Tuning model on Power Doppler US images and a 0.0415 increase from the Domain-Adaptive Tuning model. Furthermore, the mean testing quadratic weighted kappa statistic was only 0.7095. This was only a 0.0989 increase from the Vanilla Fine-Tuning model and a 0.0982 increase from the Domain-Adaptive Tuning model.

	Quadratic Weighted Kappa Statistic
Validation	0.7136 +/- 0.0196
Testing	0.7095 +/- 0.0279

Table 3.33: The quadratic weighted kappa statistic for the validation and testing sets using the Power Doppler US images on the MTL model.

Tables 3.34, 3.35, and 3.36 display the mean and standard deviation of the top 1, top 2, and top 3 validation successes, failures, and sensitivity by grade for the B-Mode validation set, respectively. In general, grade 2 had the highest sensitivity closely followed by grade 0. Grade 1 proved to be the hardest to classify class, as it consistently had the lowest sensitivity.

	Successes	Failures	Sensitivity
Grade 0	58.80 +/- 15.03	57.20 +/- 15.03	50.69% +/- 12.95%
Grade 1	3.80 +/- 3.12	38.20 +/- 3.12	9.05% +/- 7.44%
Grade 2	88.50 +/- 20.00	58.50 +/- 20.00	60.20% +/- 13.61%
Grade 3	24.30 +/- 7.50	43.70 +/- 7.50	35.74% +/- 11.03%

Table 3.34: The validation set's Top 1 Successes, Failures, and Sensitivity per grade using the MTL model with the B-Mode US image.

Tables 3.37, 3.38, and 3.39 show the mean and standard deviation of the top 1, top 2, and top 3 successes, failures, and sensitivity by grade for the B-Mode testing set, respectively. Overall, grade 2 had the highest sensitivity followed closely by grade 0. Once again, grade 1 had the lowest sensitivity.

Tables 3.40, 3.41, and 3.42 display the mean and standard deviation of the top 1, top 2, and top 3 validation successes, failures, and sensitivity by

	Successes	Failures	Sensitivity
Grade 0	94.00 +/- 13.03	22.00 +/- 13.03	81.03% +/- 11.23%
Grade 1	12.60 +/- 8.13	29.40 +/- 8.13	30.00% +/- 19.35%
Grade 2	129.90 +/- 13.63	17.10 +/- 13.63	88.37% +/- 9.27%
Grade 3	43.90 +/- 9.67	24.10 +/- 9.67	64.56% +/- 14.22%

Table 3.35: The validation set's Top 2 Successes, Failures, and Sensitivity per grade using the MTL model with the B-Mode US images.

	Successes	Failures	Sensitivity
Grade 0	111.70 +/- 4.15	4.30 +/- 4.15	96.29% +/- 3.58%
Grade 1	34.30 +/- 4.50	7.70 +/- 4.50	81.67% +/- 10.70%
Grade 2	145.40 +/- 2.76	1.60 +/- 2.76	98.91% +/- 1.88%
Grade 3	54.90 +/- 6.55	13.10 +/- 6.55	80.74% +/- 9.63%

Table 3.36: The validation set's Top 3 Successes, Failures, and Sensitivity per grade using the MTL model with the B-Mode US images.

	Successes	Failures	Sensitivity
Grade 0	55.70 +/- 13.18	48.30 +/- 13.18	53.56% +/- 12.67%
Grade 1	4.10 +/- 3.01	34.90 +/- 3.01	10.51% +/- 7.73%
Grade 2	66.70 +/- 18.03	49.30 +/- 18.03	57.50% +/- 15.54%
Grade 3	21.30 +/- 7.01	29.70 +/- 7.01	41.76% +/- 13.75%

Table 3.37: The test set's Top 1 Successes, Failures, and Sensitivity per grade using the MTL model with the B-Mode US images.

	Successes	Failures	Sensitivity
Grade 0	88.20 +/- 8.66	15.80 +/- 8.66	84.81% +/- 8.32%
Grade 1	13.40 +/- 8.13	25.60 +/- 8.13	34.36% +/- 20.84%
Grade 2	104.00 +/- 7.40	12.00 +/- 7.40	89.66% +/- 6.38%
Grade 3	35.20 +/- 6.48	15.80 +/- 6.48	69.02 +/- 12.70%

Table 3.38: The test set's Top 2 Successes, Failures, and Sensitivity per grade using the MTL model with the B-Mode US images.

	Successes	Failures	Sensitivity
Grade 0	101.00 +/- 2.79	3.00 +/- 2.79	97.12% +/- 2.69%
Grade 1	31.00 +/- 4.63	8.00 +/- 4.63	79.49% +/- 11.86%
Grade 2	114.90 +/- 1.76	1.10 +/- 1.76	99.05% +/- 1.52%
Grade 3	43.70 +/- 3.49	7.30 +/- 3.49	85.69% +/- 6.85%

Table 3.39: The test set's Top 3 Successes, Failures, and Sensitivity per grade using the MTL model with the B-Mode US images.

grade for the Power Doppler validation set, respectively. For the top 1 and top 2 tables, grade 0 had the highest sensitivity, and grade 1 had the lowest sensitivity. However, in the top 3 table, grade 2 had the highest sensitivity, while grade 3 had the lowest sensitivity.

	Successes	Failures	Sensitivity
Grade 0	134.30 +/- 13.89	32.70 +/- 13.89	80.42% +/- 8.31%
Grade 1	3.30 +/- 2.33	41.70 +/- 2.33	7.33% +/- 5.17%
Grade 2	49.60 +/- 12.71	35.40 +/- 12.71	58.35% +/- 14.96%
Grade 3	36.30 +/- 12.47	31.70 +/- 12.47	53.38% +/- 18.33%

Table 3.40: The validation set's Top 1 Successes, Failures, and Sensitivity per grade using the MTL model with the Power Doppler US images.

	Successes	Failures	Sensitivity
Grade 0	152.70 +/- 8.85	14.30 +/- 8.85	91.44% +/- 5.30%
Grade 1	20.00 +/- 9.65	25.00 +/- 9.65	44.44% +/- 21.45%
Grade 2	71.90 +/- 9.58	13.10 +/- 9.58	84.59% +/- 11.27%
Grade 3	54.80 +/- 5.72	13.20 +/- 5.72	80.59% +/- 8.42%

Table 3.41: The validation set's Top 2 Successes, Failures, and Sensitivity per grade using the MTL model with the Power Doppler US images.

Tables 3.43, 3.44, and 3.45 show the mean and standard deviation of the top 1, top 2, and top 3 successes, failures, and sensitivity by grade for the Power Doppler testing set, respectively. For the top 1 and top 2 tables, grade 0 had the highest sensitivity, and grade 1 had the lowest sensitivity. However, in the top 3 table, grade 2 had the highest sensitivity, while grade 3 had the lowest sensitivity.

	Successes	Failures	Sensitivity
Grade 0	162.60 +/- 5.82	4.40 +/- 5.82	97.37% +/- 3.48%
Grade 1	42.90 +/- 1.76	2.10 +/- 1.76	95.33% +/- 3.91%
Grade 2	84.20 +/- 1.78	0.80 +/- 1.78	99.06% +/- 2.09%
Grade 3	57.80 +/- 4.21	10.20 +/- 4.21	85.00% +/- 6.20%

Table 3.42: The validation set's Top 3 Successes, Failures, and Sensitivity per grade using the MTL model with the Power Doppler US images.

	Successes	Failures	Sensitivity
Grade 0	124.80 +/- 15.68	28.20 +/- 15.68	81.57% +/- 10.25%
Grade 1	3.50 +/- 1.96	33.50 +/- 1.96	9.46% +/- 5.3%
Grade 2	32.60 +/- 11.32	32.40 +/- 11.32	50.15% +/- 17.41%
Grade 3	26.80 +/- 8.63	22.20 +/- 8.63	54.69% +/- 17.62%

Table 3.43: The test set's Top 1 Successes, Failures, and Sensitivity per grade using the MTL model with the Power Doppler US images.

	Successes	Failures	Sensitivity
Grade 0	138.30 +/- 9.39	14.70 +/- 9.39	90.39% +/- 6.14%
Grade 1	19.20 +/- 7.81	17.80 +/- 7.81	51.89% +/- 21.10%
Grade 2	53.70 +/- 7.16	11.30 +/- 7.16	82.62% +/- 11.01%
Grade 3	37.50 +/- 4.30	11.50 +/- 4.30	76.53% +/- 8.77%

Table 3.44: The test set's Top 2 Successes, Failures, and Sensitivity per grade using the MTL model with the Power Doppler US images.

	Successes	Failures	Sensitivity
Grade 0	146.90 +/- 5.97	6.10 +/- 5.97	96.01% +/- 3.90%
Grade 1	35.60 +/- 1.91	1.40 +/- 1.91	96.22% +/- 5.16%
Grade 2	64.40 +/- 1.20	0.60 +/- 1.20	99.08% +/- 1.85%
Grade 3	42.30 +/- 4.05	6.70 +/- 4.05	86.33% +/- 8.27%

Table 3.45: The test set's Top 3 Successes, Failures, and Sensitivity per grade using the MTL model with the Power Doppler US images.

3.5 Multitask Transfer Learning with Regularization

Since B-Mode and Power Doppler US images are very similar in structure, the probability distributions of the B-Net and D-Net should behave similarly, as well.

3.5.1 Theory

Using the similar probability distribution logic, the MTL model was regularized with the Maximum Mean Discrepancies (MMD) based on borrowed ideas from the theory of Domain Adaptation Long et al. (2015). The formula for MMD is shown in Equation 3.2, where $\mathscr{H}_k$ is the reproducing kernel Hilbert space, ϕ is the feature map, $k(x^s, x^t) = \langle \phi(x^s), \phi(x^t) \rangle$.

$$d_k^2(p,q) \triangleq ||\mathbb{E}_p[\phi(x^s)] - \mathbb{E}_q[\phi(x^t)]||^2_{\mathscr{H}_k} \tag{3.2}$$

In order to implement the MMD into the code, it had to be rewritten with the use of the kernel trick as shown in Equation 3.3, where $x^s, x'^s \overset{iid}{\sim} p, x^t, x'^t \overset{iid}{\sim} q$, and $k \in K$.

$$d_k^2(p,q) = \mathbb{E}_{x^s x'^s} k(x^s, x'^s) + E_{x^t x'^t} k(x^t, x'^t) - 2E_{x^s x^t} k(x^s, x^t) \tag{3.3}$$

The MMD was added to the MTL loss function with a scalar of λ, a hyperparameter, which can be tuned to control the hardness of the regularization. The updated loss function for the MTL model is shown in Equation 3.4, where $Loss_B$ is the Cross-Entropy Loss of the B-Mode US images, $Loss_D$ is the Cross-Entropy Loss of the Power Doppler US images, and $Loss_M$ is the MMD. The validation set was used to tune the number of epochs (up to 100) and the λ values (iterated through 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3, 0.4, and 0.5).

$$Loss = 0.5 * Loss_B + 0.5 * Loss_D + \lambda * Loss_M \tag{3.4}$$

3.5.2 Results

Table 3.46 shows the best λ value (the λ value that resulted in the highest validation accuracy) per trial run. The mean λ value was 0.165, while the median was 0.1. The difference between the mean and median shows that the λ values were skewed, which can also be seen graphically in Figure 3.2.

Trial Run	λ
1	0.40
2	0.30
3	0.30
4	0.20
5	0.08
6	0.10
7	0.10
8	0.06
9	0.01
10	0.10

Table 3.46: Best λ value per trial run.

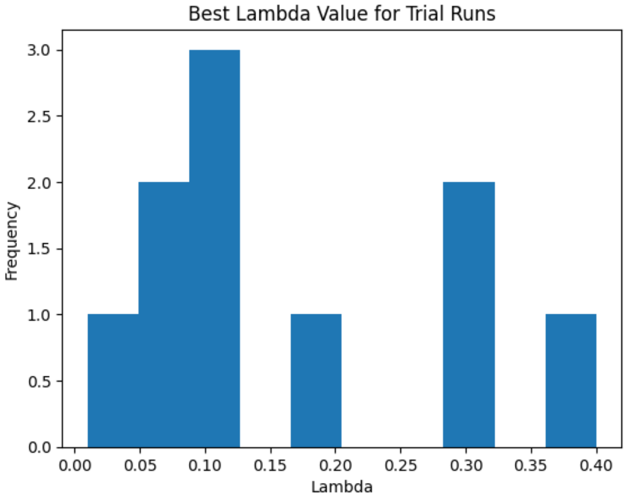


Figure 3.2: Histogram of the best λ values over the ten trial runs.

For the B-Mode images, the mean top 1 validation success rate was 52.04%, which was a 5.02% increase from the unregularized MTL model ($\lambda = 0$). Furthermore, there was an increase of 2.85% in the mean top 2 validation success rate and an increase of 0.59% in the mean top 3 validation success rate from the unregularized MTL model 3.47.

	Validation Success Rate
Top 1	52.04% +/- 1.57%
Top 2	78.02% +/- 1.29%
Top 3	93.43% +/- 1.17%

Table 3.47: Mean and standard deviation of the validation top 1, top 2, and top 3 success rates for ten trial runs using the B-Mode US images.

The mean top 1 testing success rate for the B-Mode images was 51.55%, which was an increase of 3.87% from the unregularized MTL model. The mean top 2 testing success rate was 80.52%, an increase of 2.84% from the unregularized MTL model. In addition, the mean top 3 testing success rate resulted in a 0.78% increase from the unregularized MTL model. Therefore, this model was the most satisfactory model developed in this research.

	Testing Success Rate
Top 1	51.55% +/- 2.20%
Top 2	80.52% +/- 2.17%
Top 3	94.52% +/- 1.16%

Table 3.48: Mean and standard deviation of the testing top 1, top 2, and top 3 success rates for ten trial runs using the B-Mode US images on the regularized MTL model.

As seen in Table 3.30, the mean top 1 validation success rate for the Power Doppler US images was 61.18%, which was a slight decrease from the unregularized MTL model. Furthermore, the mean top 2 validation success rate was a 0.52% decrease from the mean top 2 validation success rate of the unregularized MTL model, whereas the mean top 3 validation success rate was the same for both MTL models.

For the Power Doppler testing set, the mean top 1 success rate was 61.18% (Table 3.31). Once again, this was a slight decrease from the unregularized MTL model. The mean top 2 testing success rate was slightly higher than the unregularized model's mean top 2 success rate, but the mean top 3 success rate was slightly less. Based on the validation and testing top 1, top 2, and top

	Validation Success Rate
Top 1	61.18% +/- 3.20%
Top 2	81.51% +/- 2.23%
Top 3	95.21% +/- 1.18%

Table 3.49: Mean and standard deviation of the validation top 1, top 2, and top 3 success rates for ten trial runs using the Power Doppler US images on the regularized MTL model.

3 mean success rates, the regularized MTL model did not show significant improvement from the unregularized MTL model with respect to the Power Doppler US images.

	Testing Success Rate
Top 1	61.18% +/- 4.15%
Top 2	82.50% +/- 1.81%
Top 3	94.38% +/- 1.29%

Table 3.50: Mean and standard deviation of the testing top 1, top 2, and top 3 success rates for ten trial runs using the Power Doppler US images on the regularized MTL model.

The mean quadratic weighted kappa statistic for the B-Mode validation set was 0.4558 (Table 3.51), which was a 0.1002 increase from the unregularized MTL model. Furthermore, the mean quadratic weighted kappa statistic for the B-Mode testing set was 0.4732, which was a 0.0428 increase from the unregularized MTL model. Both of the kappa statistics show an improvement from the unregularized MTL model with respect to the B-Mode US images.

	Quadratic Weighted Kappa Statistic
Validation	0.4558 +/- 0.0390
Testing	0.4732 +/- 0.550

Table 3.51: The quadratic weighted kappa statistic for the validation and testing sets using the B-Mode US images on the regularized MTL model.

Table 3.52 displays the mean and standard deviation of the quadratic weighted kappa statistics for the Power Doppler validation and testing sets. The validation set had a mean quadratic weighted kappa statistic of 0.6756, which was a 0.0380 decrease from the unregularized MTL model. Additionally, the testing set had a mean quadratic weighted kappa statistic of 0.6581, which was a 0.0514 decrease from the unregularized MTL model. Since there was no improvement in the kappa statistics, the regularized MTL model was not considered a more powerful model with respect to the Power Doppler US images.

	Quadratic Weighted Kappa Statistic
Validation	0.6756 +/- 0.0350
Testing	0.6581 +/- 0.0466

Table 3.52: The quadratic weighted kappa statistic for the validation and testing sets using the Power Doppler US images on the regularized MTL model.

Tables 3.53, 3.54, and 3.55 display the top 1, top 2, and top 3 validation mean and standard deviation of the successes, failures, and sensitivity by grade for the B-Mode validation set, respectively. Generally, grade 2 had the highest sensitivity, followed by grade 0. For top 1 and top 2, grade 1 had the lowest sensitivity.

	Successes	Failures	Sensitivity
Grade 0	69.10 +/- 12.74	46.90 +/- 12.74	59.57% +/- 10.98%
Grade 1	2.30 +/- 1.85	39.70 +/- 1.85	5.48% +/- 4.40%
Grade 2	97.70 +/- 15.75	49.30 +/- 15.75	66.46% +/- 10.71%
Grade 3	25.00 +/- 5.22	43.00 +/- 5.22	36.76% +/- 7.67%

Table 3.53: The validation set's Top 1 Successes, Failures, and Sensitivity per grade using the regularized MTL model with the B-Mode US images.

	Successes	Failures	Sensitivity
Grade 0	102.70 +/- 7.66	13.30 +/- 7.66	88.53% +/- 6.60%
Grade 1	7.40 +/- 4.52	34.60 +/- 4.52	17.62% +/- 10.76%
Grade 2	141.40 +/- 5.99	5.60 +/- 5.99	96.19% +/- 4.07%
Grade 3	39.50 +/- 4.76	28.50 +/- 4.76	58.09% +/- 7.00%

Table 3.54: The validation set's Top 2 Successes, Failures, and Sensitivity per grade using the regularized MTL model with the B-Mode US images.

Tables 3.56, 3.57, and 3.58 show the mean and standard deviation of the top 1, top 2, and top 3 successes, failures, and sensitivity by grade for the B-Mode testing set, respectively. Overall, grade 2 has the highest sensitivity, followed by grade 0.

	Successes	Failures	Sensitivity
Grade 0	113.80 +/- 2.52	2.20 +/- 2.52	98.10% +/- 2.17%
Grade 1	33.70 +/- 6.31	8.30 +/- 6.31	80.24% +/- 15.02%
Grade 2	146.60 +/- 0.66	0.40 +/- 0.66	99.73% +/- 0.45%
Grade 3	54.40 +/- 5.02	13.60 +/- 5.02	80.00% +/- 7.39%

Table 3.55: The validation set's Top 3 Successes, Failures, and Sensitivity per grade using the regularized MTL model with the B-Mode US images.

	Successes	Failures	Sensitivity
Grade 0	61.30 +/- 11.05	42.70 +/- 11.05	58.94% +/- 10.62%
Grade 1	3.50 +/- 2.87	35.50 +/- 2.87	8.97% +/- 7.36%
Grade 2	71.30 +/- 12.57	44.70 +/- 12.57	61.47% +/- 10.84%
Grade 3	23.70 +/- 4.05	27.30 +/- 4.05	46.47% +/- 7.94%

Table 3.56: The test set's Top 1 Successes, Failures, and Sensitivity per grade using the regularized MTL model with the B-Mode US images.

	Successes	Failures	Sensitivity
Grade 0	92.40 +/- 6.58	11.60 +/- 6.58	88.85% +/- 6.32%
Grade 1	9.60 +/- 6.12	29.40 +/- 6.12	24.62% +/- 15.69%
Grade 2	111.10 +/- 5.75	4.90 +/- 5.75	95.78% +/- 4.96%
Grade 3	36.50 +/- 4.90	14.50 +/- 4.90	71.57% +/- 9.62%

Table 3.57: The test set's Top 2 Successes, Failures, and Sensitivity per grade using the regularized MTL model with the B-Mode US images.

	Successes	Failures	Sensitivity
Grade 0	101.60 +/- 3.20	2.40 +/- 3.20	97.69% +/- 3.08%
Grade 1	33.20 +/- 4.62	5.80 +/- 4.62	85.13% +/- 11.85%
Grade 2	115.70 +/- 0.64	0.30 +/- 0.64	99.74% +/- 0.55%
Grade 3	42.50 +/- 4.15	8.50 +/- 4.15	83.33% +/- 8.14%

Table 3.58: The test set's Top 3 Successes, Failures, and Sensitivity per grade using the regularized MTL model with the B-Mode US images.

Tables 3.59, 3.60, and 3.61 display the mean and standard deviation of the top 1, top 2, and top 3 validation successes, failures, and sensitivity by grade for the Power Doppler validation set, respectively. For top 1, grade 3 had the highest sensitivity, followed by grade 0. However, for the top 2 and top 3, grade 2 had the highest sensitivity, followed by grade 0.

	Successes	Failures	Sensitivity
Grade 0	121.90 +/- 16.52	45.10 +/- 16.52	72.99% +/- 9.89%
Grade 1	2.70 +/- 2.28	42.30 +/- 2.28	6.00% +/- 5.07%
Grade 2	47.60 +/- 7.34	37.40 +/- 7.34	56.00% +/- 8.63%
Grade 3	51.10 +/- 3.30	16.90 +/- 3.30	75.15% +/- 4.85%

Table 3.59: The validation set's Top 1 Successes, Failures, and Sensitivity per grade using the regularized MTL model with the Power Doppler US images.

	Successes	Failures	Sensitivity
Grade 0	147.60 +/- 14.14	19.40 +/- 14.14	88.38% +/- 8.47%
Grade 1	10.00 +/- 5.50	35.00 +/- 5.50	22.22% +/- 12.21%
Grade 2	80.00 +/- 4.96	5.00 +/- 4.96	94.12% +/- 5.84%
Grade 3	59.90 +/- 2.55	8.10 +/- 2.55	88.09% +/- 3.75%

Table 3.60: The validation set's Top 2 Successes, Failures, and Sensitivity per grade using the regularized MTL model with the Power Doppler US images.

	Successes	Failures	Sensitivity
Grade 0	161.60 +/- 3.95	5.40 +/- 3.95	96.77% +/- 2.37%
Grade 1	37.10 +/- 4.95	7.90 +/- 4.95	82.44% +/- 11.00%
Grade 2	84.80 +/- 0.60	0.20 +/- 0.60	99.76% +/- 0.71%
Grade 3	64.00 +/- 1.34	4.00 +/- 1.34	94.12% +/- 1.97%

Table 3.61: The validation set's Top 3 Successes, Failures, and Sensitivity per grade using the regularized MTL model with the Power Doppler US images.

Tables 3.62, 3.63, and 3.64 show the mean and standard deviation of the top 1, top 2, and top 3 successes, failures, and sensitivity by grade for the Power Doppler testing set, respectively. For the top 1, grade 0 had the highest mean sensitivity, followed by grade 3. However, in the top 2 and top 3 tables, grade 2 had the highest mean sensitivity, followed by grade 0. The top 2 grade 2 sensitivity topped the grade 0 sensitivity, unlike the Vanilla Fine-Tuning, Domain Adaptive Tuning, and unregularized MTL models. This could be due to the stricter regularization to behave similarly to the B-Net. The B-Mode images predicted grade 2 better potentially due to the fact that grade 2 is the largest B-Mode class, and the strict regularization forced a heavier grade 2 prediction on the D-Net.

	Successes	Failures	Sensitivity
Grade 0	114.00 +/- 15.63	39.00 +/- 15.63	74.51% +/- 10.21%
Grade 1	3.60 +/- 2.65	33.40 +/- 2.65	9.73% +/- 7.17%
Grade 2	37.00 +/- 4.65	28.00 +/- 4.65	56.92% +/- 7.15%
Grade 3	31.40 +/- 2.5	17.60 +/- 2.50	64.08% +/- 5.10%

Table 3.62: The test set's Top 1 Successes, Failures, and Sensitivity per grade using the regularized MTL model with the Power Doppler US images.

	Successes	Failures	Sensitivity
Grade 0	137.40 +/- 11.13	15.60 +/- 11.13	89.80% +/- 7.27%
Grade 1	10.50 +/- 5.14	26.50 +/- 5.14	28.38% +/- 13.90%
Grade 2	61.70 +/- 4.50	3.30 +/- 4.50	94.92% +/- 6.92%
Grade 3	41.20 +/- 3.12	7.80 +/- 3.12	84.08% +/- 6.38%

Table 3.63: The test set's Top 2 Successes, Failures, and Sensitivity per grade using the regularized MTL model with the Power Doppler US images.

	Successes	Failures	Sensitivity
Grade 0	147.40 +/- 3.32	5.60 +/- 3.32	96.34% +/- 2.17%
Grade 1	28.90 +/- 4.74	8.10 +/- 4.74	78.11% +/- 12.82%
Grade 2	64.90 +/- 0.30	0.10 +/- 0.30	99.85% +/- 0.46%
Grade 3	45.70 +/- 2.57	3.30 +/- 2.57	93.27% +/- 5.25%

Table 3.64: The test set's Top 3 Successes, Failures, and Sensitivity per grade using the regularized MTL model with the Power Doppler US images.

Chapter 4

Conclusion

Overall, the Power Doppler US images had higher success rates than the B-Mode US images. This was unsurprising since the color seen in the Power Doppler US images correlates to the sonographic RA synovitis grades Ranganath et al. (2022). Furthermore, the deep-learning models performed better than the classic machine learning models. The Domain-Adaptive Tuning model and the MTL models showed that the B-Mode US images did not provide significant improvement in predicting sonographic RA synovitis grades of Power Doppler US images. However, the MTL models proved that the information shared from the Power Doppler US images improved the prediction of the B-Mode US images.

Since the machine learning models were unsatisfactory, only the deep learning models were analyzed for potential user cases. Two main user cases were evaluated: users who wanted the model to give them the RA synovitis grade only and users who wanted more information. In other words, one model evaluation focused on a decision-making tool, whereas the other evaluation focused on a decision-assisting tool.

For the decision-making tool, the top 1 testing success rate was used to determine the best model because the user would only look at the top predicted grade in this use case. Table 4.1 shows the mean and standard deviation of the Top 1 testing success rates for each of the deep learning models discussed in Chapter 3.

With B-Mode images, there was a clear improvement from the Vanilla Finetuning model to the unregularized MTL model (mean testing success rate increase of 6.65%) and from the unregularized MTL model to the regularized MTL model (mean testing success rate increase of 3.87%). Therefore, the regularized MTL model was the optimal model for predicting sonographic RA synovitis grades for B-Mode US images.

With Power Doppler images, the mean top 1 testing success rate stayed

Model	B-Mode	Power Doppler
Vanilla Finetuning	41.03% +/- 1.62%	61.78% +/- 1.69%
Domain-Adaptive Tuning	N/A	61.68% +/- 1.57%
MTL	47.68% +/- 1.92%	61.74% +/- 3.11%
MTL + Regularization	51.55% +/- 2.20%	61.18% +/- 4.15%

Table 4.1: Top 1 testing success rates of each deep learning model type and US image mode.

relatively the same. The Vanilla Fine-Tuning model technically had the highest mean top 1 testing success rate. Plus, the Vanilla Fine-Tuning model was the simplest deep-learning model. Based on Occam's razor, the Vanilla Fine-Tuning model would be the best model since there is less than a 1% mean success rate difference between the best model (Vanilla Fine-Tuning) and the worst model (regularized MTL) with respect to the Power Doppler US images.

Therefore, the best model for a user that only uses B-Mode US images would be the regularized MTL model. However, if the user only used Power Doppler US images, the Power Doppler Vanilla Fine-Tuning model would be the best-suited model. Lastly, if the user deals with both B-Mode and Power Doppler US images, then the regularized MTL model would be the best model to limit the necessary memory space for two models due to the versatility of the regularized MTL model.

For the decision-assisting tool, the mean top 2 testing success rate was used to determine the best model for predicting sonographic RA synovitis grades. For a decision-assisting tool, the program would give the user the top 2 classes and their respective probabilities, allowing the user more information regarding potential borderline cases and to make a final decision based on the provided information instead of having the machine make the decision. The program could also give a maximum score. In other words, there would be more flexibility with the information given to the user with respect to the top 2 grades. Table 4.2 shows the mean and standard deviation of the Top 2 testing success rates for each of the deep learning models discussed in Chapter 3.

Once again, with the B-Mode US images, there was clear improvement from the Vanilla Fine-Tuning model to the unregularized MTL model (mean testing success rate improvement of 5.10%) and from the unregularized MTL model to the regularized MTL model (mean testing success rate improvement of 2.84%). Therefore, the best model to predict the sonographic RA synovitis grade of a B-Mode US image was the regularized MTL model.

With the Power Doppler US images, there was, once again, not as clear of an improvement between the deep learning models as there was with the

Model	B-Mode	Doppler
Vanilla Finetuning	72.58% +/- 2.24%	80.79% +/- 1.81%
Domain-Adaptive Tuning	N/A	79.97% +/- 0.98%
MTL	77.68% +/- 2.12%	81.81% +/- 2.27%
MTL + Regularization	80.52% +/- 2.17%	82.50% +/- 1.81%

Table 4.2: Top 2 testing success rates of each deep learning model type and US image mode.

B-Mode US images. There was only a 2.53% improvement in the mean top 2 test success rate between the best and worst deep learning models evaluated here. Therefore, Occam's razor was used again to determine that the Vanilla Fine-Tuning model was the best model.

For the decision-assisting tool the regularized MTL model would be the best model if the user only used B-Mode US images due to the high top 2 testing success rate or used both B-Mode and Power Doppler US images due to memory savings and versatility with US modes. On the other hand, the Power Doppler Vanilla Fine-Tuning model would be the best model for the user who only uses Power Doppler US images based on Occam's razor.

Overall, the regularized MTL model allowed for the best results of the evaluated models by solving some of the data scarcity issues because it combined the B-Mode or a Power Doppler US images.

4.1 Future Work

As stated in Chapter 1, there is a lack of coherency in the grading of RA Momtazmanesh et al. (2022). This lack of coherency causes some uncertainty and disagreement in the grading of RA. Therefore, an uncertainty measure for the data may be beneficial in the model prediction of RA synovitis grades.

Further tuning of the MTL model could be beneficial in its RA-predicting capabilities. More hyperparameters could be introduced and trained for better fine-tuning. Additionally, further tuning of the λ values (harshness of the regularization) would be beneficial to find the optimum λ value, which could lie outside of or in between the current evaluated λ values.

As observed throughout the deep-learning models' top 1, top 2, and top 3 successes, failures, and sensitivity tables, grade 1 was especially difficult to train. The most common grade with the minimum mean sensitivity was grade 1, and the grade with the least images overall was grade 1. For these reasons, obtaining more grade 1 data would be very beneficial. Additionally, obtaining more data overall would be beneficial in reducing the small variability between trial runs of the deep learning models and producing a more generalized model.

The evaluation of saliency maps or GradCAMs on the MTL model would allow for a better understanding of what exactly the model is looking at to make its predictions. This information could provide guidance in tuning the model to focus on different areas if need be.

The MTL modes will be tested on new B-Mode and Power Doppler US images. This will allow for a deeper understanding of how the model performs on unseen data. Lastly, once the MTL model is fully optimized, the model will be expanded to other joints in the human body.

Bibliography

(). Pennstate: Statistics online courses. URL: https://online.stat.psu.edu/stat508/lesson/9/9.2/9.2.8.

(2023). URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html.

Babcock, D. S., Patriquin, H., LaFortune, M., & Dauzat, M. (1996). Power doppler sonography: basic principles and clicical applications in children. *Pediatric Radiology*, *26*, 109–115. doi:https://doi.org/10.1007/BF01372087.

Blaivas, M., Arntfield, R., & White, M. (2020). Creation and Testing of a Deep Learning Algorithm to Automatically Identify and Label Vessels, Nerves, Tendons, and Bones on Cross-sectional Point-of-Care Ultrasound Scans for Peripheral Intravenous Catheter Placement by Novices. *American Institute of Ultrasound in Medicine*, *39*, 1721–1727.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. arXiv:1512.03385.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. arXiv:1603.05027.

Long, M., Cao, Y., Wang, J., & Jordan, M. (2015). Learning transferable features with deep adaptation networks. In F. Bach, & D. Blei (Eds.), *Proceedings of the 32nd International Conference on Machine Learning* (pp. 97–105). Lille, France: PMLR volume 37 of *Proceedings of Machine Learning Research*. URL: https://proceedings.mlr.press/v37/long15.html.

McMaster, C., Bird, A., Liew, D. F. L., R.Buchanan, R., Owen, C. E., Chapman, W. W., & Pires, D. E. V. (2022). Artificial Intelligence and Deep Learning for Rheumatologists. *American College of Rheumatology*, *74*, 1893–1905.

Momtazmanesh, S., Nowroozi, A., & Rezaei, N. (2022). Artificial intelligence in rheumatoid arthritis: Current status and future perspectives: A state-of-the-art review. *Rheumatology and Therapy*, *9*, 1249–1304. doi:10.1007/s40744-022-00475-4.

Murphy, A. (2023). Grey scale imaging (ultrasound). https://radiopaedia.org/articles/grey-scale-imaging-ultrasound?lang=us.

Ortner, R., & Leitgeb, H. (2011). Mechanizing induction. In D. M. Gabbay, S. Hartmann, & J. Woods (Eds.), *Inductive Logic* (pp. 719–772). North-Holland volume 10 of *Handbook of the History of Logic*. URL: https://www.sciencedirect.com/science/article/pii/B9780444529367500185. doi:https://doi.org/10.1016/B978-0-444-52936-7.50018-5.

Ranganath, V. K., Ben-Artzi, A., Brook, J., Suliman, Y., Floegel-Shetty, A., Woodworth, T., Taylor, M., Ramrattan, L. A., Elashoff, D., & Kaeley, G. S. (2022). Optimizing reliability of real-time sonographic examination and scoring of joint synovitis in rheumatoid arthritis. *Cureus*, . doi:10.7759/cureus.31030.

Tan, P.-N., Steinbach, M., Kumar, V., & Karpatne, A. (2019). *Introduction to Data Mining*. (2nd ed.). Pearson.

Wightman, R. (2019). Pytorch image models. https://github.com/huggingface/pytorch-image-models. doi:10.5281/zenodo.4414861.

Wu, M., Wu, H., Wu, L., Cui, C., Shi, S., Xu, J., Liu, Y., & Dong, F. (2022). A deep learning classification of metacarpophalangeal joints synovial proliferation in rheumatoid arthritis by ultrasound images. *Journal of Clinical Ultrasound*, *50*. doi:https://doi.org/10.1002/jcu.23143.

Zhang, Y., & Yang, Q. (2021). A survey on multi-task learning. arXiv:1707.08114.