# Transfer learning in the era of foundational models: application to diagnosis in rheumatology

Prashant Shekhar, PhD
*Assistant Professor of Data Science*
*Department of Mathematics*
*Embry-Riddle Aeronautical University, FL, USA*
Email: shekharp@erau.edu

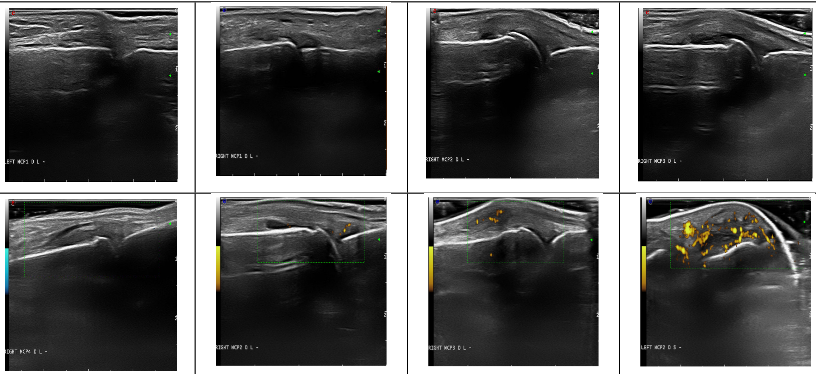
Joint work with
Dr. Gurjit Kaeley MRCP, RhMSUS, UF-COMJ
Dr. Veena Ranganath, MD, RhMSUS, UCLA
Jordan Sanders, ERAU

# Outline

# Data Source

- **B-Mode** (top): grey-scale image in which the organs and tissues of interest are depicted as points of variable brightness .
- **Power-Doppler**(bottom) is a type of ultrasound imaging that displays the strength of the Doppler signal in color.

| Grade 0 | Grade 1 | Grade 2 | Grade 3 |
|---|---|---|---|

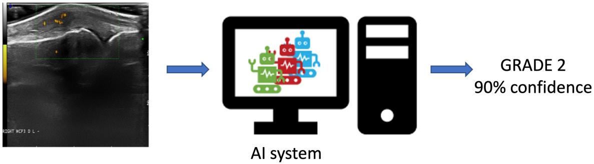


# Motivation for this work

## Problems with current synovitis grading procedures

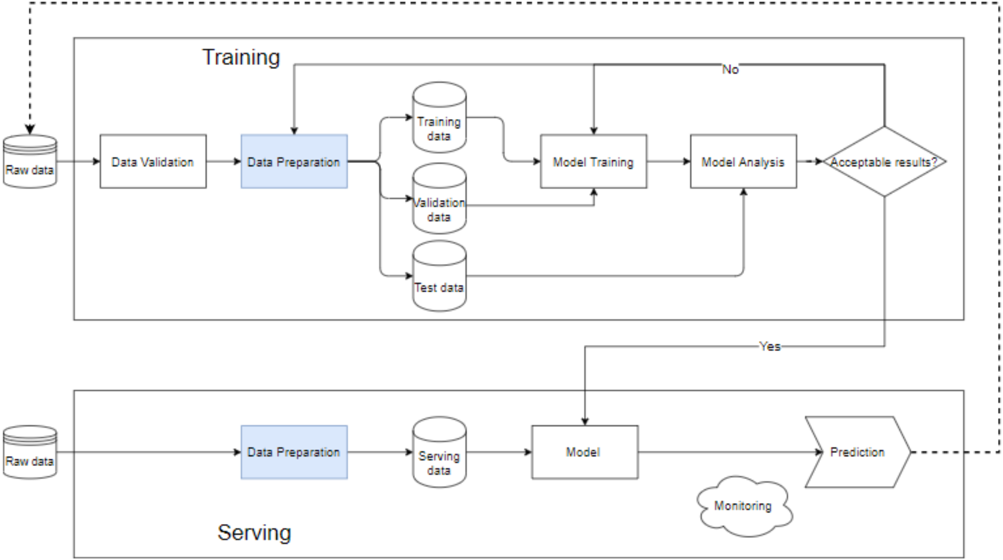
- There has been a lack of reliability in grading these images in the medical community due to a lack of universally accepted diagnostic criteria [Momtazmanesh et al., 2022]
- The human/machine variability creates an additional challenge in an efficient automated scoring system [Ranganath et al., 2022]
- There is a lack of consistency between doctors in grading these images [Momtazmanesh et al., 2022]

*Can a machine assign arthritis grades to these ultrasounds more reliably ?*

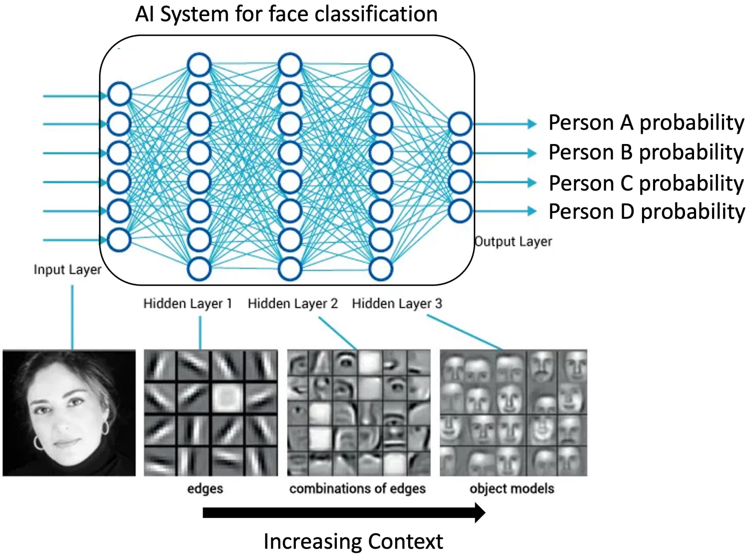


# Developing a smart AI system

# Pipeline for developing such a system



# How modern AI methods learn: a typical neural network model



## Can we directly train any such model for synovitis grading ?

1. These models are typically huge !! (millions to billions of connections/parameters)
2. They require a large number images (again of the order of 100s of thousands to millions) to find the right value of these parameters.
3. We typically don't have that many ultrasound images to learn from.

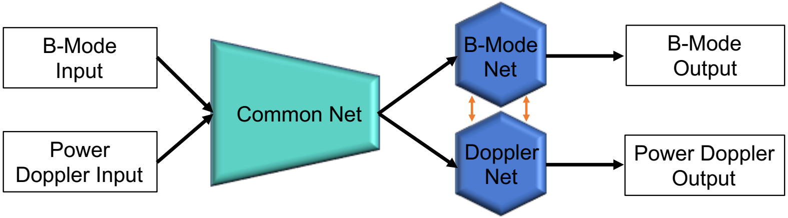
So what should we do !!

## Transfer Knowledge and Learning across models

1. **We need to move beyond the data scarcity issue**: Use models trained on publicly available large datasets and leverage their ability, to grade our ultrasound scans.
2. **We need a robust grader**: Be minimally affected by noise in data. For example different doctors, different ultrasound machines etc.
3. **We need a universal model across joints**: Be able to handle scans of different body joints.
4. **We need a universal model across ultrasound modes**: Be able to handle grading for both B-Mode and Power Doppler scans.

# Proposed AI system

# Regularized-Multitask Transfer Learning (Reg-MTL)



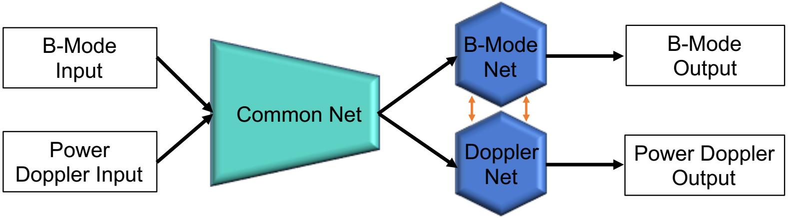

The model accepts both B-Mode and Power Doppler Mode Scans

An AI model that learns features of both modes and has been pretrained on millions of regular images

Smaller individual models of both modes with feature alignment based regularization

Depending on the mode of input, you get the corresponding grade prediction

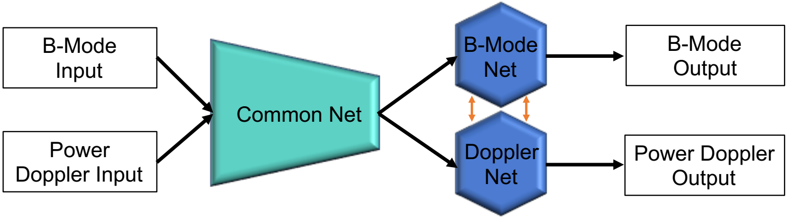
# Regularized-Multitask Transfer Learning (Reg-MTL)




## Model Inputs

- Jointly training with B-Mode and Power Doppler mode scans handles the data scarcity problem.
- Jointly training allows the model to give importance to both modes and learn from both modes.
- Since algorithm searches for similar relevant image features across both modes jointly, it learns to differentiate between noise and signal.

# Regularized-Multitask Transfer Learning (Reg-MTL)




## Common Net

- Here we use ResNetv2 [He et al., 2016] architecture having 67 million parameters.
- This model is pre-trained on ImageNet21K dataset which contains 14 million images from 21,000 classes. These includes natural images belonging to classes like cars, animals etc.
- A model pre-trained on such images has been shown to extract good features even from unrelated images such as ultrasound scans.
- This idea is referred to as **Transfer learning** [Zhang et al. 2021] in the AI community.

# Transfer Learning idea: Image Models

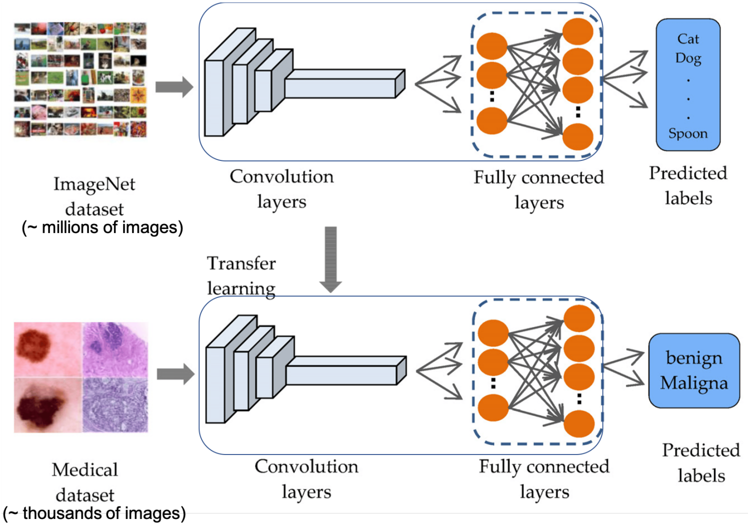
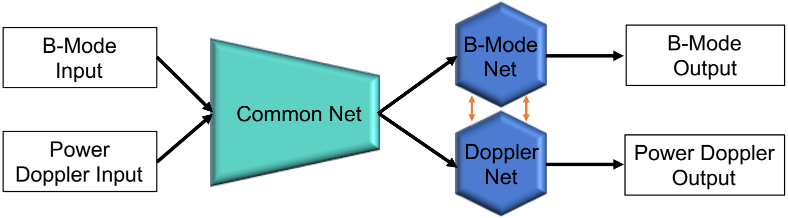



Figure: Adapted from [Mukhlif et al., 2023]

# Regularized-Multitask Transfer Learning (Reg-MTL)




## B-Mode Net and Doppler Net

- Mode specific models that use the joint features learnt by Common Net, to make grade predictions.
- These models have very few parameters (100s to 1000s). So tuning with few images is possible.
- We also constrain the behavior of these individual models to avoid learning from noise in images. This is done through feature alignment to regularize the representations of features learnt across B-Mode Net and Doppler Net.

# Feature alignment based regularization

## Maximum Mean Discrepancies for feature aligning

- Intuitively, the representations learnt across B-Mode Net and Doppler Net should be similar. This is because the causal features across both modes have similar shapes.
- We use this logic to regularize the behavior of the MTL model.
- To implement this, we borrow ideas from the theory of Domain Adaptation [Long et al., 2015] and use Maximum Mean Discrepancies (MMD) (1) to restrict the two individual nets to learn similar representation

$$d_k^2(p,q) \triangleq ||\mathbb{E}_p[\phi(x^{B-Net})] - \mathbb{E}_q[\phi(x^{D-Net})]||^2_{\mathcal{H}_k} \tag{1}$$

Where $\mathcal{H}_k$ is the reproducing kernel Hilbert space (RKHS), $\phi$ is the feature map, and $x^{B-Net}$, $x^{D-Net}$ are corresponding feature representations learnt across BMode Net and Doppler Net. The inner product in the RKHS is defined as $k(x^s, x^t) = \langle \phi(x^s), \phi(x^t) \rangle$

# Training Reg-MTL model

## Cost function for Reg-MTL model

MMD is added to the MTL loss function with a scalar $\lambda$, a hyperparameter, which can be tuned to control the hardness of the constraint.

$$Loss = 0.5 * Loss_{BM} + 0.5 * Loss_{DM} + \lambda * Loss_{MMD} \quad (2)$$

Where $Loss_{BM}$ is the Cross-Entropy Loss of the B-Mode images, $Loss_{DM}$ is the Cross-Entropy Loss of the Power Doppler images, and $Loss_{MMD}$ is the MMD loss

| $\lambda$ | *0.01* | *0.02* | *0.03* | *0.04* | *0.05* | *0.06* | *0.07* | *0.08* | *0.09* | *0.10* | *0.20* | *0.30* | *0.40* | *0.50* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **BM** | 47% | 42% | 49% | 49% | 49% | 50% | 49% | 49% | 49% | 50% | 50% | 51% | 52% | 51% |
| **DM** | 62% | 61% | 63% | 64% | 60% | 60% | 60% | 60% | 60% | 61% | 62% | 59% | 59% | 60% |

Table: Finetuning $\lambda$ for best trade-off between accuracy and robustness (validation accuracy)

# Results

# Data Distribution

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

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

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

Table: Table showing the breakdown of images by dataset.

# Fine-tuning separate ResNetv2 pretrained models for each US mode

## B-Mode

| | Validation Accuracy |
|---|---|
| Top 1 | 42.23% +/- 1.98% |
| Top 2 | 70.78% +/- 2.74% |
| Top 3 | 89.87% +/- 1.43% |

| | Testing Accuracy |
|---|---|
| Top 1 | 41.03% +/- 1.62% |
| Top 2 | 72.58% +/- 2.24% |
| Top 3 | 91.06% +/- 1.32% |

## Doppler Mode

| | Validation Accuracy |
|---|---|
| Top 1 | 63.26% +/- 1.15% |
| Top 2 | 81.15% +/- 1.49% |
| Top 3 | 94.90% +/- 1.12% |

| | Testing Accuracy |
|---|---|
| Top 1 | 61.78% +/- 1.69% |
| Top 2 | 80.79% +/- 1.81% |
| Top 3 | 94.21% +/- 1.31% |

Top 1, top 2, and top 3 accuracy mean $\pm$ standard deviation of validation/testing set for ten trial runs.

# Fine-tuning one universal model (MTL) for both US modes jointly

## B-Mode

| | Validation Accuracy |
|---|---|
| Top 1 | 52.04% +/- 1.57% |
| Top 2 | 78.02% +/- 1.29% |
| Top 3 | 93.43% +/- 1.17% |

| | Testing Accuracy |
|---|---|
| Top 1 | 51.55% +/- 2.20% |
| Top 2 | 80.52% +/- 2.17% |
| Top 3 | 94.52% +/- 1.16% |

## Doppler Mode

| | Validation Accuracy |
|---|---|
| Top 1 | 61.18% +/- 3.20% |
| Top 2 | 81.51% +/- 2.23% |
| Top 3 | 95.21% +/- 1.18% |

| | Testing Accuracy |
|---|---|
| Top 1 | 61.18% +/- 4.15% |
| Top 2 | 82.50% +/- 1.81% |
| Top 3 | 94.38% +/- 1.29% |

Top 1, top 2, and top 3 accuracy mean $\pm$ standard deviation of validation/testing set for ten trial runs.

- B-Mode is always more difficult for arthritis grade prediction (relative to Power Doppler mode).
- The accuracy numbers reported for both B-Mode and Power-Doppler mode are close to the performance level of a human rheumatologist [Ranganath et al.].

## Benefits of Multitask learning

- **Improvement in B-Mode performance**: Optimizing Reg-MTL jointly on B-Mode and Power Doppler Mode improves the performance on B-Mode drastically.
- **Improvement in Power Doppler Mode performance**: Optimizing Reg-MTL jointly on B-Mode and Power Doppler Mode improves the sensitivity of grade 2 in Power Doppler. This is because the model is able to transfer knowledge from large number of grade 2 samples in B-Mode. *Grade 2 is usually the most difficult grade to predict correctly*.

## Conclusion and Future work

- [x] Need for moving beyond data scarcity issue (Transfer Learning + Multitask Learning)
- [x] Need for a universal model across ultrasound modes (Multitask Learning)
- [ ] Need for a robust grader (Causal Machine Learning ?)
- [ ] Need for a universal model across joints (Domain Generalization ?)

# References


1. Momtazmanesh, Sara, Ali Nowroozi, and Nima Rezaei. "Artificial intelligence in rheumatoid arthritis: current status and future perspectives: a state-of-the-art review." Rheumatology and Therapy 9.5 (2022): 1249-1304.
2. Ranganath, Veena K., et al. "Optimizing Reliability of Real-Time Sonographic Examination and Scoring of Joint Synovitis in Rheumatoid Arthritis." Cureus 14.11 (2022).
3. He, Kaiming, et al. "Identity mappings in deep residual networks." Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer International Publishing, 2016.
4. Zhang, Yu, and Qiang Yang. "A survey on multi-task learning." IEEE Transactions on Knowledge and Data Engineering 34.12 (2021): 5586-5609.
5. Mukhlif, Abdulrahman Abbas, Belal Al-Khateeb, and Mazin Abed Mohammed. "Incorporating a Novel Dual Transfer Learning Approach for Medical Images." Sensors 23.2 (2023): 570.
6. Long, Mingsheng, et al. "Learning transferable features with deep adaptation networks." International conference on machine learning. PMLR, 2015