COLUMBIA UNIVERSITY
DATA SCIENCE INSTITUTE

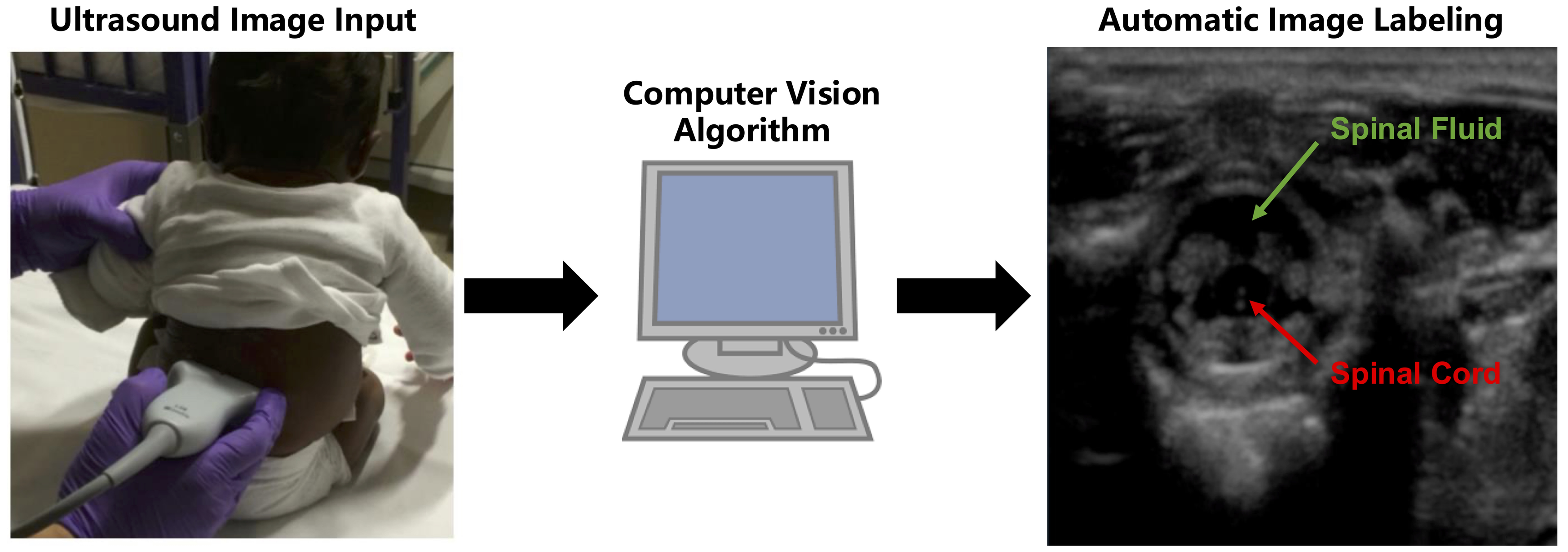
Machine Learning for Automated Ultrasound-Guided Infant Lumbar Puncture

Abhik Biswas [1] Isaac Peabody [1] Harinder Mashiana [1] Mohini Bhave [1] Shrinjay Kaushik [1]
Faculty and Industry Advisors: Dr. David Kessler [2] Dr. Elia Lima-Walton [3] Ritvik Khandelwal [2]
[1]Data Science Institute, Columbia University [2]Columbia University Irving Medical Center [3]Elsevier

COLUMBIA
COLUMBIA UNIVERSITY
IRVING MEDICAL CENTER

Abstract

This project aims to improve the success of ultrasound-guided lumbar punctures in infants by leveraging machine learning. Lumbar punctures are essential for diagnosing critical conditions like meningitis, yet their success rate in infants is remarkably low (50–60%) due to the challenge of identifying spinal landmarks [2]. While ultrasound imaging enhances visualization, its interpretation requires specialized expertise [2]. We aim to develop automated computer vision models to detect spinal cord and cerebrospinal fluid (CSF) in ultrasound images, thereby supporting physicians in performing accurate procedures. The models developed will assess image quality, detect spinal landmarks, and generalize across diverse datasets to maximize clinical applicability. By addressing this healthcare challenge, the project seeks to combine data science techniques with clinical insights to create an impactful solution.



Dataset Overview

Our dataset includes 13 spinal ultrasound videos segmented into 3,390 labeled frames (Set 1) and 681 frames (Set 2). Each frame is annotated for image quality, presence of spinal cord, and CSF. Additionally, a set of 32 spinal ultrasound images from varied sources was provided for testing model generalization (Set 3). All three sets show notable differences, including varied contrast and intensity, indicating the need for image pre-processing and standardization.

To achieve optimal model performance on both seen and unseen ultrasound images, we propose the following dataset split. Models will be initially developed on a combination of Set 1 and Set 2, using a traditional train-test split. The exact breakdown of Set 1 and Set 2 is a hyperparameter that can be tuned. To assess model performance on unseen images, an additional test phase will occur using Set 3, which is not included in the initial development or test set.

Preprocessing

To enhance model generalization, we incorporated a number of on-the-fly augmentation techniques during training, including random horizontal and vertical flips, random rotation, cropping (4 types - crop the left and right most sides of the image, crop from all sides, crop the top and bottom, and no cropping) at random and contrast adjustment. Of these techniques, cropping seemed the most promising, so we focused on tuning the exact cropping specifications including type of cropping, percent of images cropped, and amount of original image cropped. We also applied normalization to the original image.

Classification Models - ResNet18 & Vision Transformers (ViT)

We approached our problem as a multi-label classification task, predicting image quality, presence of spinal cord, and presence of CSF simultaneously. The primary models tested for this task were **ResNet18** and **Vision Transformers** (ViT) [3][4]:

- ResNet18 [3]:
  - **18-layer convolutional neural network** with residual connections, enabling deeper training
  - Approximately **11.7 million parameters**

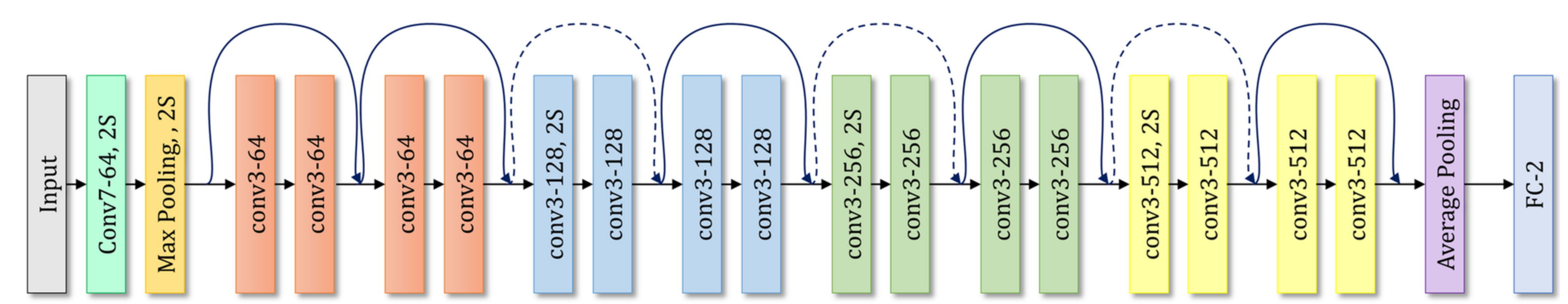


Figure 1. ResNet18 Model Architecture

- Vision Transformers (ViT) [1]:
  - Leverage **self-attention** mechanisms to process images as sequences of patches, capturing global dependencies, **approximately 86 million parameters**
  - Can be **fine-tuned with LoRA adapters** and mixed precision training to address computational constraints, reduced to **approximately 2.58 million parameters**

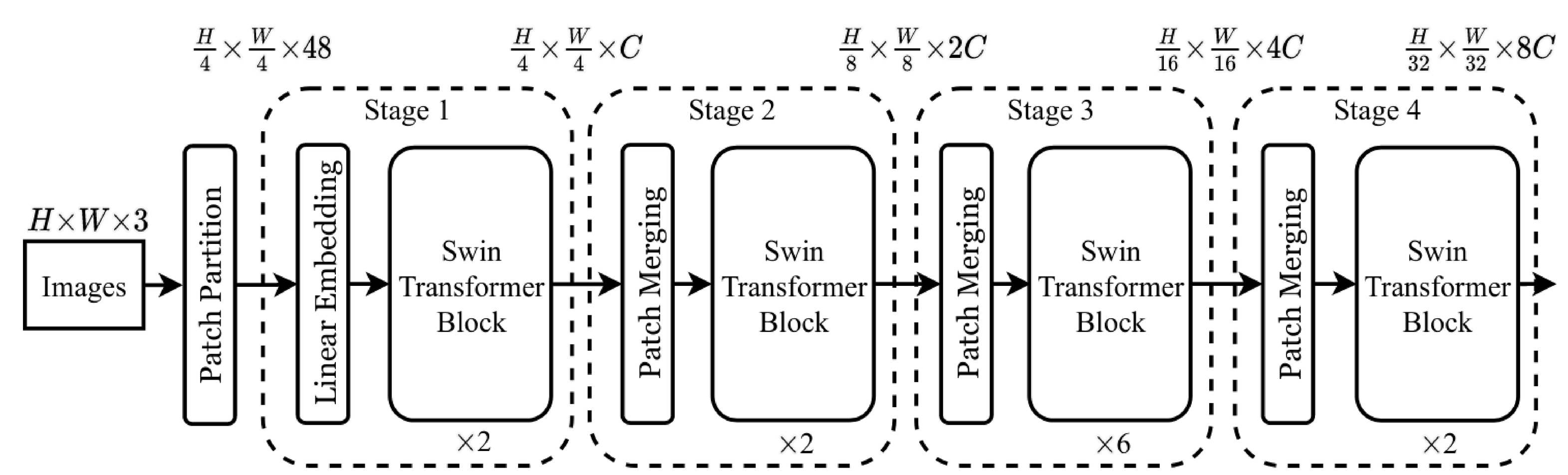


Figure 2. ViT Model Architecture

ResNet - Results

For **ResNet18**, we tuned both cropping percentage and train-val-test split of Set 2 images. The performance on the unseen set of images (Set 3) are below:

Model Performance Metrics				
Metrics	Accuracy	Precision	Recall	F1-score
Norm (70-10-20 Split)	0.7813	0.8431	0.7679	0.8037
Norm+Crop(0.9) (70-10-20 Split)	0.7579	0.8200	0.7455	0.7810
Norm+Crop(0.8) (70-10-20 Split)	**0.8958**	0.9259	0.8929	0.9091
Norm+Crop(0.8) (60-10-30 Split)	0.8021	0.9302	0.7143	0.8081
Norm+Crop(0.8) (50-10-40 Split)	**0.8958**	0.9107	**0.9107**	**0.9107**
Norm+Crop(0.8) (40-10-50 Split)	0.8646	**0.9574**	0.8036	0.8738
Norm+Crop(0.8) (30-10-60 Split)	0.7500	0.8810	0.6607	0.7551
Norm+Crop(0.7) (70-10-20 Split)	0.8125	0.9318	0.7321	0.8200

Table 1: Summary of ResNet18 Model Performance.

Results - Vision Transformers

Model Performance Metrics				
Model + Adaptations	Accuracy	Jaccard	Hamming	F1-score
ViT - Base:Norm, LoRA: rank= 32	0.9787	0.9525	0.011	0.9855
ViT - Pretrained (Thyroid USG), Norm+Crop(0.8) LoRA: rank= 16	**0.9869**	**0.9683**	**0.0072**	**0.9909**
DeiT - Pretrained (ImageNet-1k): Norm+Crop(0.8) LoRA: rank= 32	0.9771	0.9520	0.0131	0.9821

Table 2: Summary of ViT Model Performance.

Misclassifications

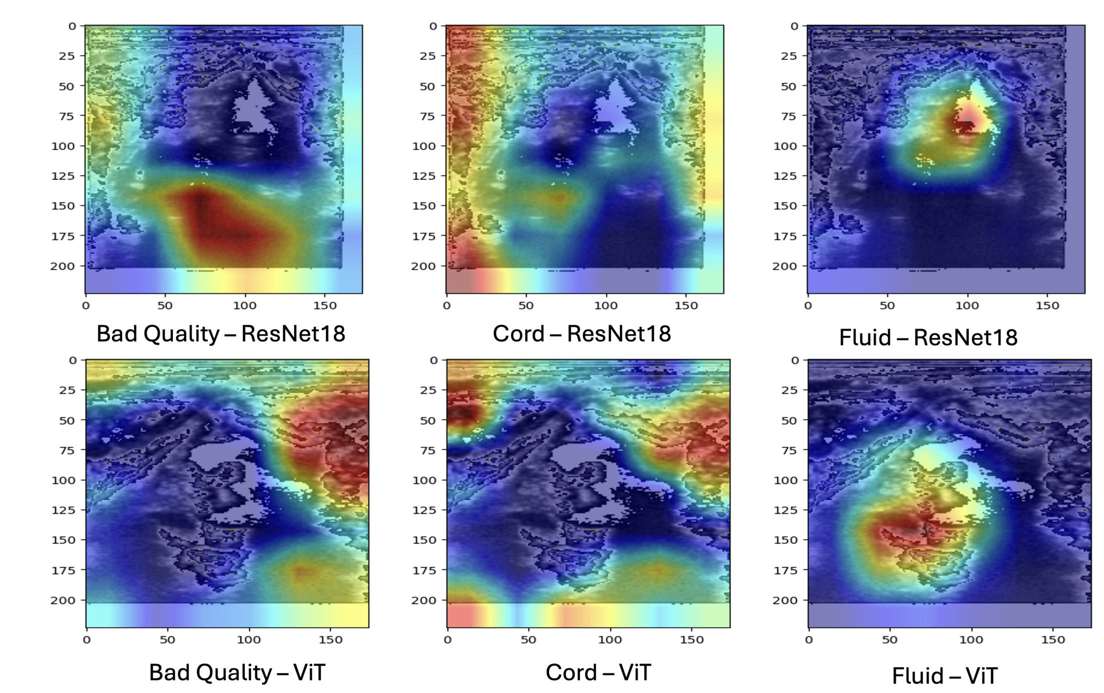


Figure 3. Misclassifications - these Gradient-weighted Color Activation Maps show the regions of an image which led the model to give incorrect predictions.

Conclusions and Future Work

ViT and ResNet18 effectively capture patterns in ultrasound videos, with ResNet18 excelling on static datasets and ViT showing potential with further optimization and larger datasets.

Future work includes refining data augmentation, optimizing temporal models, and ensuring robustness through cross-validation to enhance spinal landmark detection and improve lumbar puncture accuracy in infants, with more training data to further enhance model performance.

References

[1] Alexey Dosovitskiy et al.
An image is worth 16x16 words: Transformers for image recognition at scale.
CoRR, abs/2010.11929, 2020.

[2] David Kessler et al.
Preprocedural ultrasound for infant lumbar puncture: A randomized clinical trial.
Academic Emergency Medicine, 25(9):1027–1034, Sep 2018.
Epub 2018 May 16.

[3] Kaiming He et al.
Deep residual learning for image recognition.
CoRR, abs/1512.03385, 2015.