Detecting Lumbar Spine Degenerative Conditions From MRI Data

Aaron Schwall[1], Shawn Xu[1], Marion Bauman[1]

Georgetown University[1]

June 4, 2024

1 Introduction

According to the World Health Organization, lower back (lumbar spine) pain is a leading cause of disability across the world, affecting more than 600 million people (2023). Many people will experience lumbar spine pain at some point in their lives, and it is increasingly prevalent with age. Efficient and accurate diagnosis of lumbar spine conditions is essential for dictating treatment and speeding rehabilitation (Richards et al., 2024). The importance of this project is to use machine learning and artificial intelligence in tandem with computer vision techniques to aid in the detection as well as the classification of five specific lumbar spine degenerative conditions. The five degenerative conditions being studied are: Left neural Foraminal Narrowing, Right Neural Foraminal Narrowing, Left Subarticular Stenosis, Right Subarticular Stenosis, and Spinal Canal Stenosis. Rapid detection of degenerative conditions through machine learning will lead to improved medical treatment, improving patient outcomes and quality of life for millions worldwide. According to current research, there has already been work done in the Deep Learning space where Convolutional Neural Networks have been used for detection and classification of medical conditions. For our specific use case, there has not been research done on the previously mentioned 5 conditions that we hope to detect and classify. Due to the current lack of an automated method, we hope to evaluate multiple Neural Network models that are accurate enough to detect and classify the 5 degenerative conditions.

2 Background and Literature Review

There have been many studies in the past looking at the use of computer vision techniques in diagnosing and treating lower back pain based on MRI scans. D'Antoni, et al. found 76 articles in the PubMed database relating to diagnosing and treating lower back pain using computer vision (D'Antoni et al., 2021). They found that the most commonly used applications of computer vision were feature detection and segmentation, and that in recent years the use of deep learning models has been much greater than the use of traditional image processing techniques.

In general, the majority of recent studies on applying computer vision and deep learning to MRI scans focus on the use of Convolutional Neural Networks (CNNs). For example, Liawrungrueang, et al. used CNNs to identify Intervertebral disc degeneration (IDD). They found that when using T-2 weighted MRI scans their CNN model was able to classify and grade spinal injuries using the Pfirrmann grading system with an accuracy of more than 95 percent (Liawrungrueang et al., 2023). Another study used T-2 weighted images to train CNNs to detect and label various lumbar spine degenerative changes (Lehnen et al., 2021). The study attempted to detect five different types of degenerative changes including disk herniation and canal stenosis. They were able to detect these changes with varying accuracies, with the highest accuracy being 98 percent for detecting canal stenosis. The lowest accuracy was 76 percent for detecting disk bulgings.

What we have not seen in studies are applications of CNNs and other machine learning techniques to the particular conditions that are addressed in our dataset. While there are many studies examining various lumbar spine conditions, the novelty of our dataset provides a new area to explore the application of CNNs and other advanced machine learning techniques to MRI scans.

3 Data Gathering and Research Questions

In order to develop a model that can diagnose lumbar spine degenerative conditions, we selected a dataset of approximately 48,000 thousand MRIs. Our dataset is sourced from Kaggle, and by conducting this research project, we are participating in an active competition hosted on Kaggle by the Radiological Society of North America and the American Society of Neuroradiology to solve this detection and multi-class classification problem (Richards et al., 2024). The data was curated by compiling lumbar spine MRI scans and their diagnoses from universities across the globe including MRI images from the United States of America, Thailand, Australia, Turkey, and more. Over 50 medical experts contributed to the dataset by reviewing and labeling the MRI scans with the diagnosis, severity, and relevant vertebrae. This global compilation is to ensure generalization in solutions.

Through the usage of CNNs and other advanced machine learning techniques, image-based models can be trained to distinguish between different spinal conditions displayed in MRI scans. We posit these research questions:

- **Research Question 1 (RQ1):** How effectively can computer vision algorithms diagnose lumbar spine degenerative conditions based on MRI scans?
- **Research Question 2 (RQ2):** How effectively can computer vision algorithms classify the severity of lumbar spine degenerative conditions based on MRI scans?
- **Research Question 3 (RQ3):** How effectively can computer vision algorithms locate the relevant vertebrae for lumbar spine degenerative conditions?

Through these research questions, we aim to understand the capabilities of computer vision models for diagnosing lumbar spine conditions. Alongside a disorder diagnosis, we aim to determine the power of machine learning for classifying the severity of the disorder and the location of the degeneration in the vertebrae. Through enhanced machine learning detection and granular diagnosis of lumbar spine conditions, medical teams can efficiently diagnose and pinpoint spinal degeneration, guiding treatment to alleviate pain. Effective detection of lumbar spine conditions in MRIs can lead to rapid response and improved intervention to combat the leading cause of disability worldwide.

4 Data Description

The data obtained from the Kaggle competition by the Radiological Society of North America (RSNA) was sourced from eight institutions across five continents. The purpose of this multi-institutional curated dataset is to create a ground truth dataset that will enable generalization for the development of artificial intelligence tools in lumbar spine disease classification.

The dataset consists primarily of MRI images of the lumbar spine region saved in the .DCM file format. The .DCM format is short for DICOM (Digital Imaging and Communication in medicine), which is the generally accepted format for medical imagery. The images are organized by study, with each study having a unique study ID. The studies are then further broken down by series of imagery. Most of the studies in our dataset contain three imaging series, however some contain more. Within each series are the individual MRI images taken during that series. There are also four CSV files included in the data. Train.csv contains the labels for the training data. The variables in the file are study_id and the label, which has the condition as well as the level of severity. An example label would be spinal_canal_stenosis_l1_l2_severe. In this example spinal canal stenosis is the condition, l1 and l2 are the vertebrae where the condition is occurring, and severe is the level of severity. The train_label_coordinates.csv file contains the study_id, the series_id, the instance_number, the condition, level, and the coordinates of the center of the area where the condition occurs in the image. The study_id, condition, and level are the same as with the training labels. The series_id and the instance number provide the exact image where the condition can be found, and the coordinates point out where we can find the condition in the image. The last CSV files included are the [train/test]_series_description.csv files. These files contain the study id, the series id, and the orientation of the scan. Examples of the scan orientation are Axial T2, and Sagittal T1. The orientation of the scan has two parts. First is the plane on which the image is oriented. The second is the type of weighting used in the scan, primarily T1 or T2. The type of weighting affects what types of tissues are brighter in the scan. Fatty tissue and the inner parts of bones show up brighter in T1 images, while the watery inner spine appears brighter in T2 images (Suri et al., 2023).

Our data contains approximately 48,000 individual images. Due to the nature of the kaggle competition, the test data provided is not labeled. In order to adequately train our models we will split the provided training data into training and testing sets. Given that the study sizes are not uniform, we will divide our data at the study level. 80 percent of the studies will be randomly allocated to the training set while 15 will become the test set. We will then form a validation set using the remaining 5 percent.

4.1 Dependent Variable

The primary purpose of this project is to identify the medical condition affecting the lumbar spine. Thus, the dependent variable for this project is the lumbar spine condition that we are attempting to classify. The five conditions listed by the RSNA presented in the MRI images are Left neural Foraminal Narrowing, Right Neural Foraminal Narrowing, Left Subarticular Stenosis, Right Subarticular Stenosis, and Spinal Canal Stenosis.

Although the Kaggle competition specified the classification of severity and vertebral level in the final predictions, for our project scope we plan to ignore both and focus only on the given degenerative condition labels. This will allow us to focus on achieving the highest classification accuracy possible in a short period of time.

Figure 1

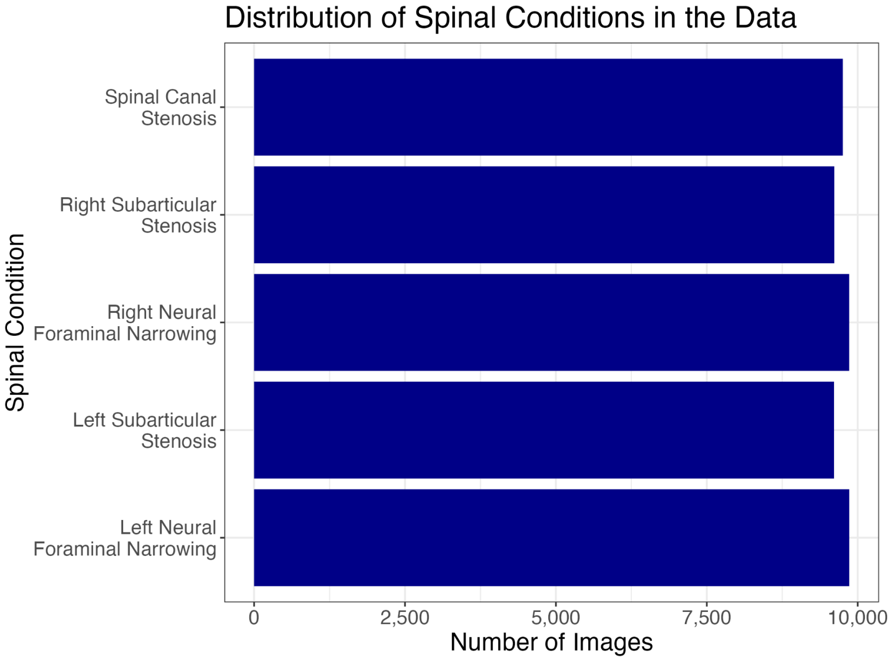

Figure 1: Each of the five spinal conditions has a little less than 10,000 samples for use in training the models.

4.2 Data Preprocessing Methods

The import process for our massive pool of DICOM image data will consist of a standardized pipeline for all DICOM MRI images. This way we treat each image similarly to eliminate any biases in our degenerative condition classification task. We plan to follow a similar approach as detailed in an example notebook provided in the Kaggle competition (Suri et al., 2023). The approach consists of reading in each DICOM image with the Pydicom module and normalizing pixel values with OpenCV.

With how specific the lumbar spine conditions are, we plan to perform preprocessing steps that will zoom in on key areas of lumbar spine conditions. These key areas are labeled with X and Y coordinates in the train_label_coordinates.csv file. Our team hypothesizes that cropping these images to specify on these labeled areas relevant to the degenerative conditions will eliminate a large amount of image noise. With our decision to attempt cropping to eliminate image noise, we also plan to standardize our image dimensions to be the same so that the cropped images are able to be fed into our chosen Neural Network model. The dimensions will be decided later but we hope to downsize the dimensions for shorter training times.

Figure 2

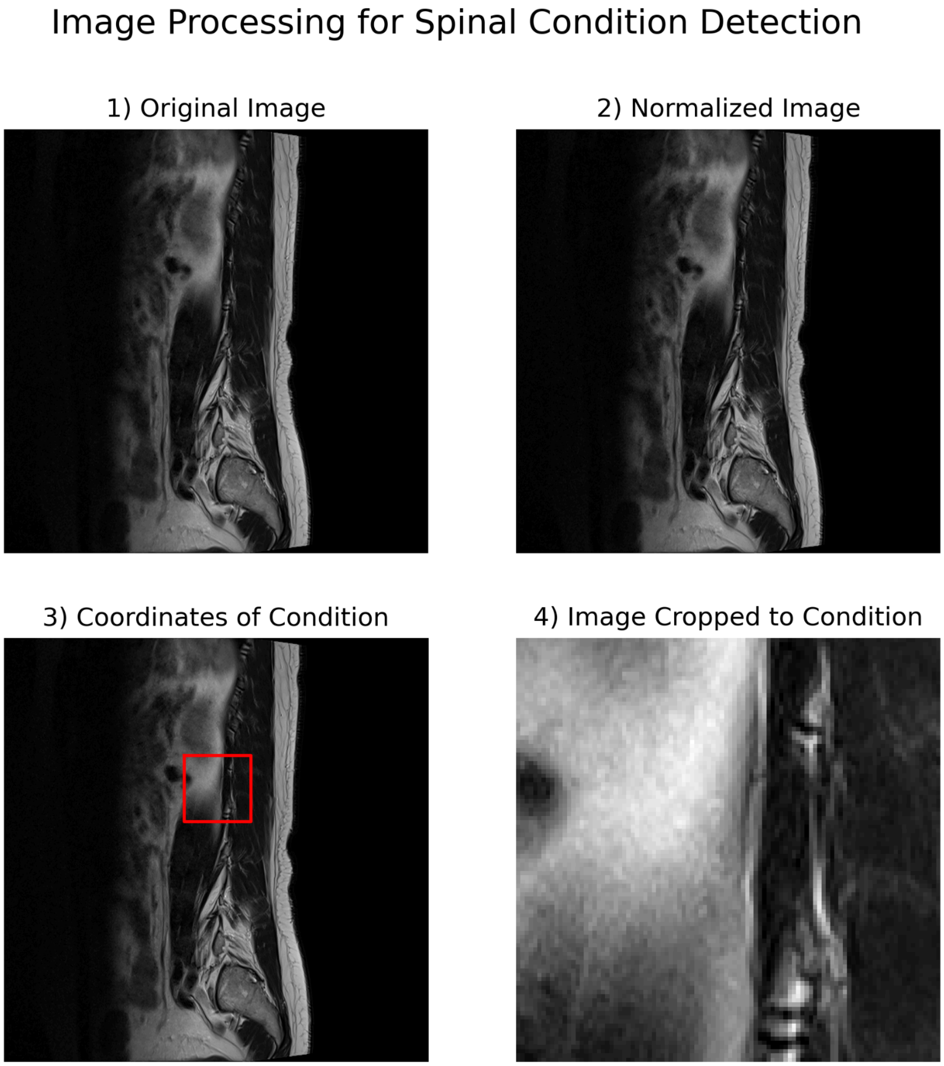

Image Processing for Spinal Condition Detection

1) Original Image

2) Normalized Image

3) Coordinates of Condition

4) Image Cropped to Condition

Figure 2: The image preprocessing pipeline for our modeling normalizes the images, identifies the coordinates of the spinal condition, and crops the image to focus on the condition.

4.3 Training and Test Set Construction

We were given training and test DICOM image data by the RSNA from the Kaggle competition. However, we were not given metadata regarding the labeling of the test dataset. We thus turned our attention to splitting the training dataset to create our training and test datasets. As mentioned before, since the study sizes were not uniformly distributed, we will split the dataset study-wise. The training dataset will contain 80 percent of the studies, 15 percent will form the test dataset, and the remaining 5 percent will be for the validation dataset.

With the validation dataset, we plan to utilize it to hypertune the parameters and hyperparameters of our chosen Neural Network model. As done in "NTIRE 2020 Challenge on Spectral Reconstruction From an RGB Image", we will compare the same model architectures with different hyperparameters and select the model that scores the best (Arad et al., 2020). To evaluate which model performs the classification task the best, we will employ metrics like Top-1 Accuracy. We will also visualize the models' performance with graphical methods like confusion matrices.

4.4 Deep Learning Methods

To carry out our analysis we will test various types of CNN (convolutional neural nets) based models, as these models have been shown in the past to be highly effective for analyzing imagery. The six models that we will look into are ResNet-101, ResNet-152, VGG-11, VGG-16, MNASNet-0.75, and MNASNet-1.0. We will analyze the results of these different models to determine which one can be best applied to our MRI image classification problem.

Transfer learning could also be utilized if time is a major constraint or if we are limited on data (Zhang et al., 2022). If needed, we hope to tune popular pre-trained models with state-of-the-art performance for our classification purposes. To do this we would look into popular pre-trained Neural Network models in the computer vision space. These models would then be trained on our DICOM image dataset until convergence to adjust the parameters to be accustomed to our lumbar spine condition classification task.

5 Analysis

With our training, validation, and test datasets created, our next step is training the chosen deep learning models on the data. To make the process efficient, we will use a combination of Early Stopping strategies and Pytorch's DataLoader function. We will also attempt to perform the multiprocessing steps with a GPU (graphics processing unit) instead of a CPU for quicker training times.

For the image data being fed into our training model, we plan on adjusting our pipeline to crop the original DICOM images of size 640 x 640 pixels into an image of about 100 x 100 pixels in size. The image will be cropped around the labeled condition. This will eliminate a lot of image noise that does not pertain to lumbar spine conditions as well as speed up the training process as fewer pixels need to be processed per image.

As mentioned above, we will apply a CNN based model for our computer vision classification task as CNN models have proven to be well performing models in this space. Our model will consist of multiple sequences of layers based on specific CNN architectures. The first layer will take in image data and output an intermediate feature map. These will then be fed into the next layer and so on. These initial layers that produce intermediate feature maps utilize convolutional kernels to convolve with the input images through matrix dot products. The selected model will then evaluate these extracted feature maps by predicting the condition labeled in the given image. This evaluation will employ Categorical Cross Entropy Loss as the primary loss function to compare the predicted label and the ground truth label. This comparison gives each label a probability that sums up to 1.0 where the highest probability label will be outputted as the predicted label. Categorical Cross Entropy Loss is also very suitable for our project task as we are attempting to do classification on multiple features. That is, the five lumbar spine degenerative conditions.

For our project goal we are trying to maximize the likelihood of predicting the correct condition for our input images. To maximize this likelihood, we plan on testing multiple optimization algorithms like Adam, RMSProp, and SGD with momentum. The best optimization method will be selected as part of our hyper-parameter tuning process.

To determine how well the final trained and tuned model performs on unseen data, we will feed in the test dataset into the model and calculate performance metrics based on its predicted labels. This will primarily consist of the Top-1 Accuracy in lumbar spine degenerative condition classification for each model.

5.1 Experiment

Since our goal is to develop a Deep Learning model that would be able to predict 5 specific degenerative lumbar spine conditions given MRI data, we thus would like to perform four main tasks:

a. Extracting Features of the neural network from the given image data.
b. Training our model to learn from these extracted feature maps for each degenerative condition.
c. Evaluating our models' performance in accurately detecting lumbar spine condition.
d. What lumbar spine condition is detected in given unseen image data?

We outline below the model architectures that we will evaluate on our MRI dataset. In Table 1, we describe the 4 different model architecture names, the number of layers, and a short description for each.

Table 1 Deep Learning Architectures

Architecture	**No. of layers**	**Short description**
VGG	11, 16	An architecture that is deeper (i.e., has more layers of neurons) and obtains better performance than AlexNet by using effective 3 × 3 convolutional filters (Norouzzadeh, et al.).
MNASNet	0.75, 1.0	This architecture is computationally efficient (using 12 times fewer parameters than AlexNet) while offering high accuracy (Tan, et al.).
ResNet	101, 152	The winning architecture of the 2016 ImageNet competition (Norouzzadeh, et al.). The number of layers for the ResNet architecture can be different. We try 101 and 152 layers in this work.

5.2 Results

To get a generalized and accurate performance comparison of the models of interest, we compare each model's performance on unseen Test data. The resulting Test accuracy for each model is shown in Table 2.

Table 2 Test Accuracy of different models on task (d)

Architecture	Top-1 accuracy, %
VGG11	91.4
VGG16	93.4
MNASNet-0.75	87.3
MNASNet-1.0	88.2
ResNet-101	94.6
ResNet-152	94.4

5.2.1 Accuracy Performance

To visually compare the test accuracy of each model that we have evaluated, we illustrated them in the following barplot.

Figure 3

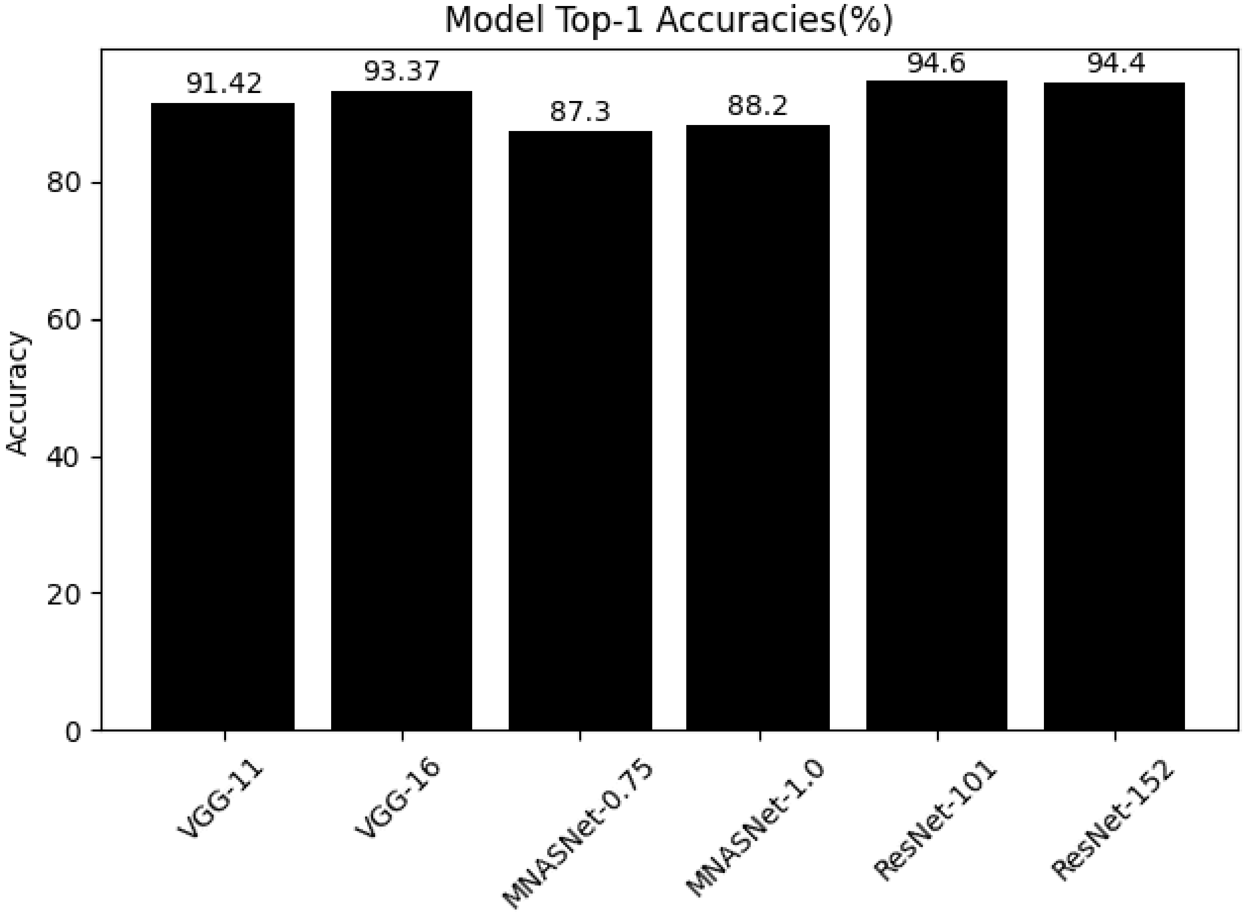

Figure 3: A comparison in the performance of each model evaluated. The bar plot utilizes the test accuracies of each model to compare each model's performance.

With the use of transfer learning, we were able to witness strong accuracies from each model we have evaluated. The deeper model architectures provide a somewhat stronger performance compared to the more shallow models.

5.2.2 Training and Test Loss Performance

As an additional step to compare the model architectures, we looked at training and test losses of each model per epoch. We view the loss versus epoch in two line plots shown below. Note that each model has a different number of epochs run due to the use of early stopping in model training.

Figure 4

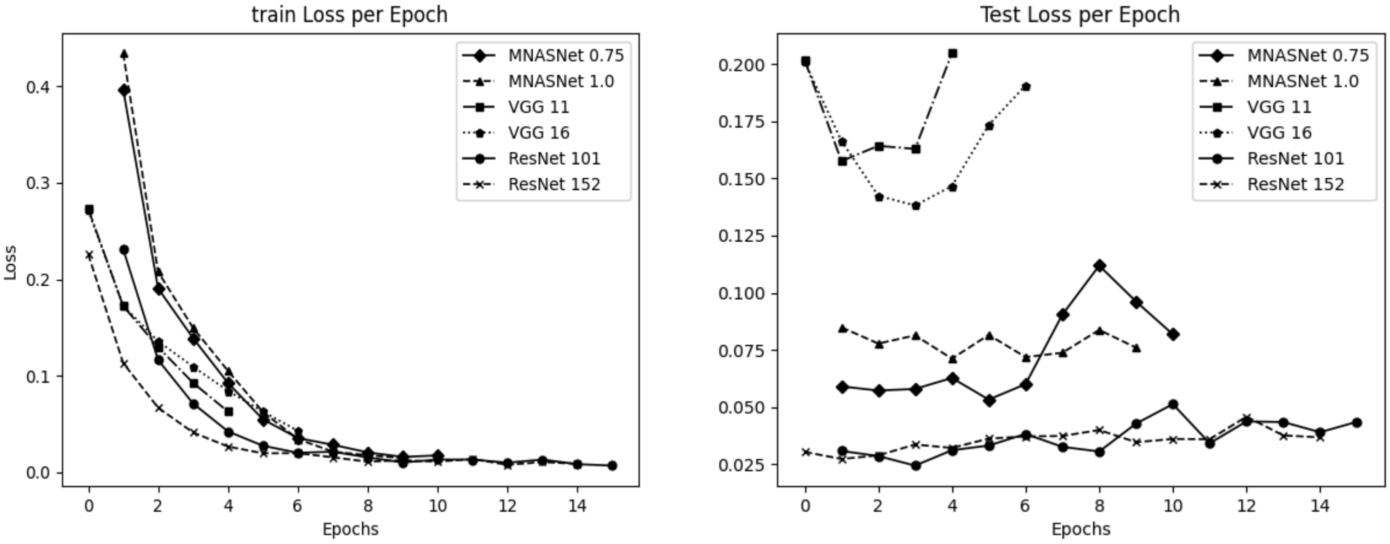

Figure 4: A side-by-side view of the Train (left) and Test (right) Loss per Epoch line plots.

There seemed to be some overfitting from the Train dataset with the Test dataset having a stagnant or increase in loss per epoch. This is especially apparent in our ResNet models where the Test Loss per Epoch, on average, is increasing. That being the case, overfitting should not be an issue in the ResNet architecture as Residual Blocks are used to remove these issues. A more extensive investigation might be needed to discover the cause of this increasing loss trend. Notably, while both ResNet and VGG have similar training loss minimization, ResNet has a much lower Test loss than VGG.

5.2.3 Test Prediction Performance

The following confusion matrices were evaluated with each model on our test dataset to visually interpret the performance of each model on each lumbar spine condition classification task.

Figure 5

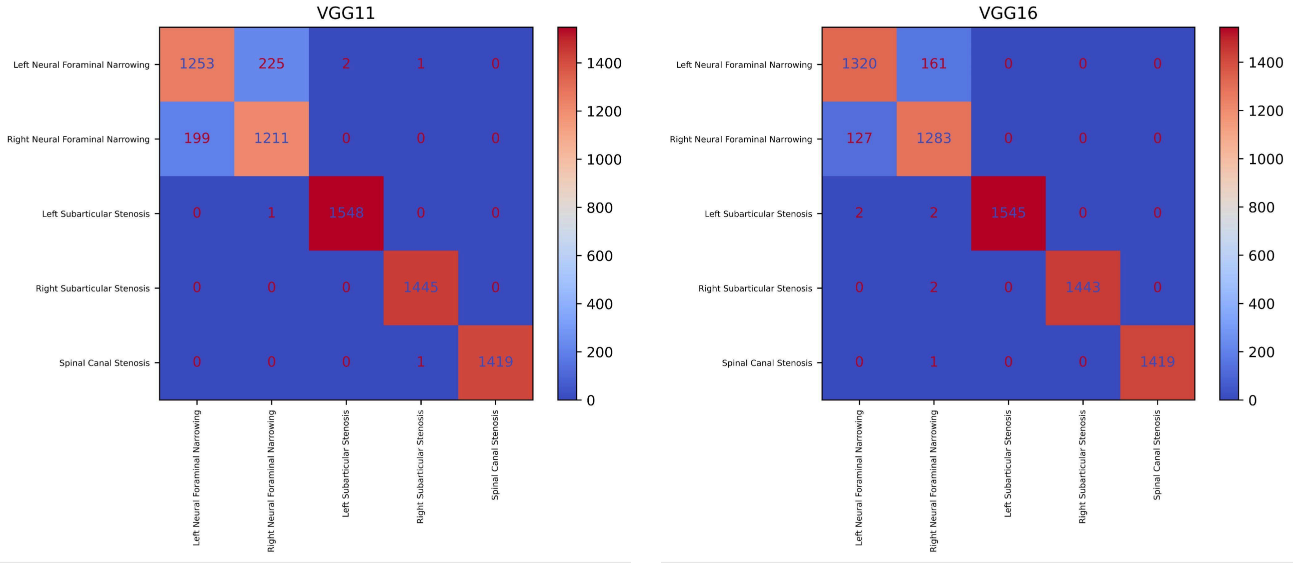

Figure 5: A side-by-side view of confusion matrices to illustrate VGG model architectures' performance in predicting Lumbar Spine Degenerative Conditions.

Figure 6

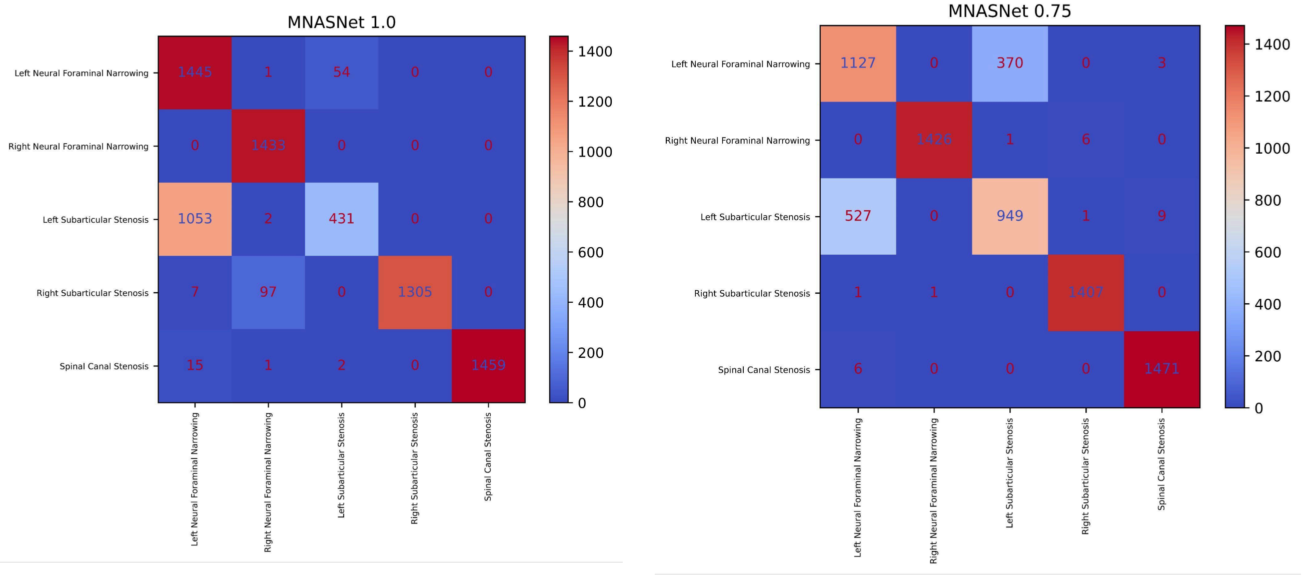

Figure 6: A side-by-side view of confusion matrices to illustrate MNASNet model architectures' performance in predicting Lumbar Spine Degenerative Conditions.

Figure 7

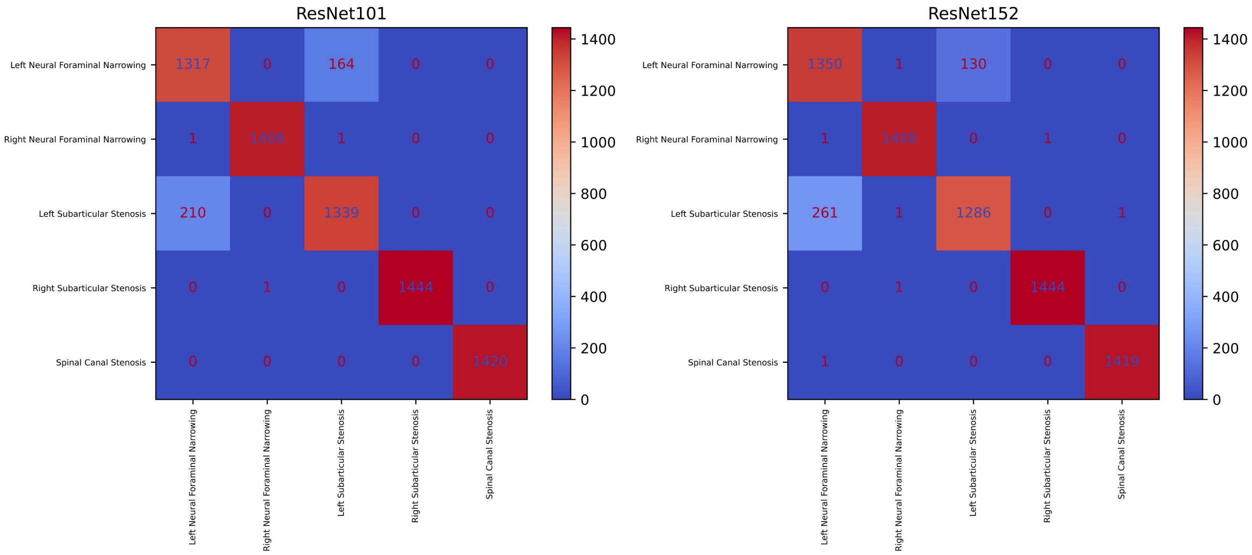

Figure 7: A side-by-side view of confusion matrices to illustrate ResNet model architectures' performance in predicting Lumbar Spine Degenerative Conditions.

We can see when looking at the confusion matrices that the different models of the same type have similar error patterns, mixing up two similar classes. With VGG, the model struggles to differentiate between Left and Right Neural Foraminal Narrowing. For ResNet, the most common error is between Left Neural Foraminal Narrowing and Left Subarticular Stenosis. Despite these errors, this indicates that the different depth levels of the model types are learning the same overall patterns within the data resulting in strong prediction performances.

6 Conclusion

The purpose of this study was to determine if we could predict 5 specific degenerative lumbar spine conditions with the given MRI data using deep learning models. To do this, we tested two different versions of each of three popular CNN based models and compared the results. We found that of the three, MNASNet 0.75 and 1 performed the worst with test accuracies of 87.3% and 88.2%. VGG 11 and 16 performed slightly better, achieving test accuracies of 91.42% and 93.37% respectively. We found that the best models were ResNet 101 and 152, achieving 94.6% and 94.4% accuracies respectively. Despite achieving these high accuracies, more work needs to be done in the future to study how these models perform on the full uncropped images. Overall we are satisfied with the strong prediction performance achieved in this study, and believe that these results can be useful to medical practitioners attempting to speed up the initial detection time of degenerative lumbar spine conditions.

References

Arad, B., Timofte, R., Ben-Shahar, O., Lin, Y. T., & Finlayson, G. D. (2020). Ntire 2020 challenge on spectral reconstruction from an rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 446-447). https://doi.org/10.48550/arXiv.2005.03412

D'Antoni F, Russo F, Ambrosio L, Vollero L, Vadalà G, Merone M, Papalia R, Denaro V. Artificial Intelligence and Computer Vision in Low Back Pain: A Systematic Review. Int J Environ Res Public Health. 2021 Oct 17;18(20):10909. doi: 10.3390/ijerph182010909. PMID: 34682647; PMCID: PMC8535895. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8535895/

Lehnen NC, Haase R, Faber J, Rüber T, Vatter H, Radbruch A, Schmeel FC. Detection of Degenerative Changes on MR Images of the Lumbar Spine with a Convolutional Neural Network: A Feasibility Study. *Diagnostics*. 2021; 11(5):902. https://doi.org/10.3390/diagnostics11050902.

Liawrungrueang W, Kim P, Kotheeranurak V, Jitpakdee K, Sarasombath P. Automatic Detection, Classification, and Grading of Lumbar Intervertebral Disc Degeneration Using an Artificial Neural Network Model. *Diagnostics*. 2023; 13(4):663. https://doi.org/10.3390/diagnostics13040663

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., & Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with Deep Learning. Proceedings of the National Academy of Sciences, 115(25). https://doi.org/10.1073/pnas.1719367115

Richards, T., Talbott, J., Ball, R., Colak, E., Flanders, A., Kitamura, F., Mongan, J., Prevedello, L., & Vazirabad, M. (2024). *RSNA 2024 Lumbar Spine Degenerative Classification*. Kaggle.

https://kaggle.com/competitions/rsna-2024-lumbar-spine-degenerative-classification

Saleem S, Aslam HM, Rehmani MA, Raees A, Alvi AA, Ashraf J. Lumbar disc degenerative disease: disc degeneration symptoms and magnetic resonance image findings. Asian Spine J. 2013 Dec;7(4):322-34. doi: 10.4184/asj.2013.7.4.322. Epub 2013 Nov 28. PMID: 24353850; PMCID: PMC3863659. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3863659/

Suri, A., Wentland, A., & Trivedi, H. (2023). *Anatomy & image visualization overview-RSNA RAIDS*. Kaggle. Retrieved from https://www.kaggle.com/code/abhinavsuri/anatomy-image-visualization-overview-rsna-raids/notebook

Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., & Le, Q. V. (2019). MnasNet: Platform-aware neural architecture search for mobile. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2019.00293

World Health Organization (2023, June 19). *Low back pain*. Retrieved June 4, 2024, from https://www.who.int/news-room/fact-sheets/detail/low-back-pain

Zhang, S., Lee, D., Singh, P. V., & Srinivasan, K. (2022). What makes a good image? airbnb demand analytics leveraging interpretable image features. Management Science, 68(8), 5644–5666. https://doi.org/10.1287/mnsc.2021.4175