Software Engineering Department
Braude College

Capstone Project Phase A – 61998

Lumbar Spine Degenerative Classification using an Optimized CNN
(24-2-R-1)

Supervisor:

Miri Weiss-Cohen

Submitters:

Almog Kadosh 315699439

Daniel Zada 318223278

Table Of Content

1.Introduction

Low back pain ranks as the primary cause of disability globally, affecting 619 million individuals in 2020, as reported by the World Health Organization. This common affliction often arises from degenerative spinal disorders like spondylosis, which deteriorates intervertebral discs and leads to spinal stenosis and nerve compression. LBP affects individuals of all ages, peaking at 50-55 years and being more prevalent in women. Chronic LBP significantly impacts older adults aged 80-85, causing work loss, reduced quality of life, and economic burden. This study evaluates standardized definitions of degenerative changes to reduce variability in lumbar spine Magnetic resonance imaging (MRI) interpretations among subspecialty-trained doctors and explores using an AI-trained model to classify lower back problems using MRI images and LBP classifications. [1]

The human back is composed of several key components. The spine, which consists of thirty-three vertebrae
(Fig. 1), is divided into four regions: 7 cervical vertebrae (C1-C7), 12 thoracic vertebrae (T1-T12), 5 lumbar vertebrae (L1-L5), and fused sacral and coccygeal vertebrae. Intervertebral discs act as cushions between the vertebrae, providing shock absorption and flexibility. Various muscles, including the erector spinae, trapezius, and latissimus dorsi, support and move the spine. Ligaments connect bones and stabilize joints, while facet joints between vertebrae allow limited movement and contribute to spinal stability. Finally, the spinal cord runs through the vertebrae, with nerve roots branching out to innervate different body parts [1].

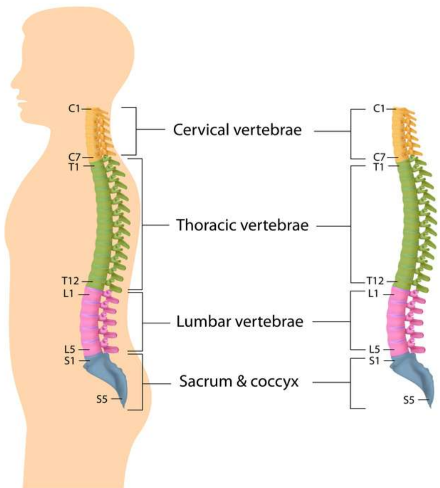


Fig.1 human spinal skeleton showing cervical (yellow), thoracic (green), lumbar (pink) and saccro-coccygeal (blue) regions [1]

Back pain can be caused by a variety of factors, including both lifestyle choices and underlying medical conditions. Risk factors such as weak core muscles, obesity, physically demanding jobs, and chronic stress can all contribute to back pain. Age, genetics, and even improper lifting techniques can also play a role. Back pain can also stem from structural problems in the spine, discs, muscles, or ligaments, such as strains, sprains, and degenerative disc disease. In some cases, inflammatory conditions like ankylosing spondylitis or other medical conditions like osteoporosis, fibromyalgia, or even kidney stones can be the culprit. While the specific cause of back pain may not always be clear, understanding these various risk factors and potential underlying conditions is a crucial first step in addressing and preventing back pain.[3]

One prevalent lumbar spine issue is Degenerative Spondylolisthesis Lumbar Spine (DSLS), primarily affecting older adults, particularly women over 50. DSLS involves the anterior displacement of one vertebra over another, typically occurring at the L4-L5 level due to degenerative changes in intervertebral discs and facet joints.

DSLS has three distinct types, differentiated by four key parameters as depict in Figure 2. Type 1 is characterized by a Pelvic Incidence (PI) - Lumbar Lordosis (LL) mismatch of less than 10 degrees, indicating a relatively aligned spinal configuration. In contrast, Type 2 is defined by a PI-LL mismatch exceeding 10 degrees, suggesting more significant spinal misalignment. The most severe, Type 3, is distinguished by a Sagittal Vertical Axis (SVA) measurement greater than 40 mm, indicating substantial sagittal imbalance. These classifications facilitate a targeted approach to understanding and treating DSLS. [4]

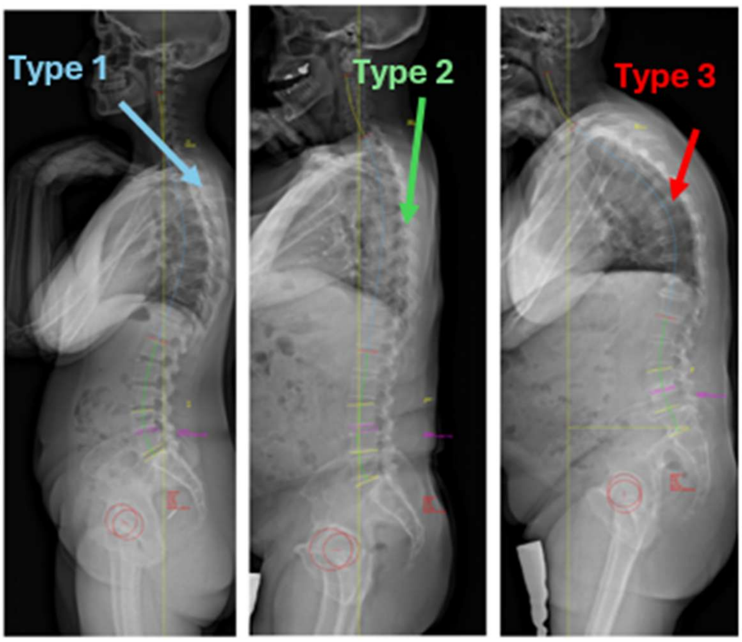


Fig. 2 Type 1(blue arrow), on lateral standing X-ray view. type 2 (green arrow), Altered global LL, compensated malalignment on X-ray view. Type 3 (red arrow), Harmonious and aligned spine on X-ray view. [4]

DSLS manifests through symptoms like lower back pain, leg pain, and neurogenic claudication, necessitating accurate imaging for effective management. advocate for a novel classification system focusing on sagittal alignment to refine diagnostic precision and tailor treatment strategies based on demographic data, radiographic features, and quality-of-life outcomes.

Further complicating lumbar spine diagnoses are congenital anomalies like lumbosacral transitional vertebrae (LSTV), discussed by Their study evaluates the efficacy of anteroposterior lumbar plain radiographs (AP-LPR) versus coronal reconstructed CT images (CT-CRIs) in classifying these vertebrae using the Castellvi classification system. The findings highlight the limitations of AP-LPR in capturing detailed anatomical nuances, advocating for the superior accuracy of CT-CRIs for precise classification [4].

MRI is a crucial tool for identifying and diagnosing back problems, as it provides detailed images of the spine's structures, including vertebrae, discs, and nerves. This capability allows for accurate detection of issues like spinal stenosis, nerve compression, and degenerative changes. That is why we focus on using MRI to figure out the precise nature of these conditions, ensuring that diagnoses are accurate and treatment plans are well-informed.

The following challenges of Using MRI scans for classifying Lumbar back are as follows:
One significant challenge, as highlighted by Miskin et al. [2], is the absence of standardized classifications for MRI images of lumbar spine degenerations. This lack of standardization leads to inconsistencies in diagnosing and assessing these conditions. Moreover, there is notable variability in interpretations both within and between intra- and inter-subspecialties, exacerbating the difficulty of achieving accurate and consistent evaluations.
For instance, the variability in interpretations is particularly evident between neuroradiologists and musculoskeletal radiologists. Although both specialties focus on the issue of lumbar spine degenerations, they approach it from distinct perspectives. Neuroradiologists often emphasize neurological implications and spinal cord involvement, while musculoskeletal radiologists may concentrate more on bone and soft tissue abnormalities. This divergence in focus can lead to differences in diagnostic criteria and treatment recommendations, further complicating the standardization of MRI assessments for these conditions.
Furthermore, the absence of standardized classification systems for spinal degeneration leads to inconsistent evaluations, which compromises the accuracy of clinical interventions. This inconsistency poses a challenge for the timely and effective management of conditions such as degenerative spondylolisthesis and congenital anomalies like lumbosacral transitional vertebrae. Without uniform criteria, the variability in diagnosis and treatment can delay appropriate care and potentially impact patient outcomes.
Standardizing diagnostic criteria is crucial for ensuring consistency and accuracy in evaluating back problems. This can be achieved using artificial intelligence (AI) and convolutional neural networks (CNNs). By leveraging these advanced technologies, we can automate the analysis of imaging data, leading to more uniform and reliable diagnostic outcomes. AI and CNNs can help establish standardized classifications, reducing variability and enhancing the precision of diagnoses and treatment plans for conditions like spinal degeneration and other related disorders.
Our goal is to classify lumbar spine degenerations using CNN to improve the efficiency and accuracy of evaluating MRI images. By automating the diagnostic procedure, we aim to provide assisting medical frameworks to ease physicians' workloads. As a result, we aim to significantly reduce the diagnostic time for patients, resulting in more efficient and effective treatment plans for conditions such as spinal degeneration.

2. Literature review

Miskin et al. [2] proposed a standardized classification system for lumbar spine degeneration on magnetic resonance imaging (MRI) to bridge the interpretative gap between subspecialty-trained neuroradiologists (NR) and musculoskeletal radiologists (MSK). This initiative aimed to address the "inter-subspecialty variability" that significantly affects the consistency of MRI interpretations related to lumbar spine conditions.

The standardized definitions for four critical parameters of lumbar degeneration: spinal canal stenosis (SCS), neural foraminal stenosis (NFS) (fig. 3), lateral recess stenosis (LRS), and facet osteoarthritis (FO) were established through a multidisciplinary consensus involving orthopedic and neurological spine surgeons, physiatrists, and radiologists. The consensus aimed to make MRI readings more uniform and clinically relevant, focusing on parameters commonly targeted for therapeutic interventions.

For the study, fifty consecutive patient examinations previously utilized were reassessed. These patients had undergone MRI scans (fig. 3) for chronic low back pain, radiculopathy (refers to a set of conditions in which one or more nerves are affected and do not work properly), or clinical symptoms indicative of spinal stenosis, all without the use of intravenous gadolinium contrast (chemical substances used in MRI scans). To ensure unbiased results, a washout period of 11 months was observed before re-evaluating the scans, during which the six participating radiologists were blinded to all identifying patient data.

To enhance the radiologists' familiarity with the new classification system, comprehensive training materials were provided:

1. Schematic diagrams and representative MR images (not from the fifty cases) compiled into a Microsoft PowerPoint presentation (Redmond, WA).
2. A concise single-page summary PDF (Adobe, Mountain View, CA) featuring schematic diagrams and descriptions of the classifications.
3. A 19-minute instructional video that detailed the classification system.

These resources aimed to standardize the evaluation criteria and reduce variability in diagnosing and grading lumbar spine conditions.

Fig 3. A sagittal image at the level of the exiting left L5 nerve root (arrow) was reviewed. Before training, the six readers varied in their assessments of the degree of neural foraminal stenosis: one described it as mild, three as mild-moderate, one as moderate, and one as moderate-severe. Following training, five readers classified it as moderate, while one classified it as severe. [2]

Following this training, statistical analysis using the Cohen's Kappa coefficient demonstrated significant improvements in inter-subspecialty agreement for all evaluated parameters post-training, the agreement levels for SCS and NFS shifted from moderate to substantial, and for LRS and FO from slight to fair and moderate, respectively. This indicated a successful reduction in diagnostic variance and an increase in consistency across evaluations.
However, the study noted a limitation in the absence of a definitive gold standard for reporting lumbar spine degeneration, suggesting that future research could explore whether this more consistent classification correlates with improved patient outcomes after specific interventions.

In conclusion, the implementation of a standardized classification system, developed through extensive collaboration and supported with detailed training materials, proved effective in enhancing the consistency of lumbar spine MRI reporting between neuroradiology and musculoskeletal radiology subspecialties. This approach has potential implications for improving diagnostic accuracy and patient management in clinical settings.
Gille et al. [4] introduce a novel classification system for Degenerative Spondylolisthesis (DSLS) that categorizes the condition based on sagittal alignment, incorporating measures of pelvic incidence and lumbar lordosis. This innovative approach aims to provide a more comprehensive framework that correlates with patient symptoms, radiographic parameters, and outcomes following therapeutic interventions. The classification system divides DSLS into three main types based on the severity of sagittal misalignment and the extent of pelvic compensation, which are crucial for tailoring treatment strategies.
Previous classification systems have primarily focused on localized assessments without incorporating the broader implications of spinal and pelvic alignment. These systems often fail to provide a holistic framework necessary for optimal surgical planning and outcome prediction. They have emphasized the importance of sagittal balance in the prognosis and management of spinal disorders, advocating for its integration into DSLS classification. The introduction of a classification system that considers sagittal alignment and pelvic compensation marks a significant advancement in the understanding and treatment of DSLS, offering a more precise method to predict patient outcomes and guide therapeutic decisions.
This system categorizes DSLS into three main types, each with specific subtypes, to improve diagnostic precision and treatment planning.
The first type, known as the Harmonious Spine, is characterized by a pelvic incidence-lumbar lordosis (PI-LL) mismatch of less than 10 degrees. It includes two subtypes: 1A, which has preserved segmental lordosis (SL), and 1B, which has altered SL but preserved overall lumbar lordosis (LL). The second type, referred to as Compensated Malalignment, features a PI-LL mismatch greater than 10 degrees. It is further divided into two subtypes based on pelvic tilt (PT): 2A, with low PT, and 2B, with high PT, indicating more significant compensation. The third type, Altered Global Sagittal Alignment, involves severe malalignment with significant pelvic compensation and a high sagittal vertical axis (SVA).
The study retrospectively analyzed a group of patients treated for DSLS, utilizing a combination of health-related quality-of-life scales and radiographic parameters. This approach allowed for a detailed correlation between the new classification types and various clinical outcomes, offering a more refined understanding of DSLS pathophysiology and treatment implications.

Lisheng Hou et al. [5] have proposed a new classification system for lumbosacral transitional vertebrae (LSTV), intended to become the gold standard to counteract the significant variability and frequent misclassifications associated with current diagnostic methods. Their approach is to replace the conventional anteroposterior view of the lumbar plain radiograph (AP-LPR) with coronal reconstructed CT images (CT-CRIs), which they argue provide a more accurate and reliable method for detecting and classifying LSTVs according to the Castellvi classification.
The need for a new method arises from the limitations of the AP-LPR, which has been found to be insufficient for accurate detection and classification of morphological abnormalities associated with

LSTVs. Specifically, the study highlights that AP-LPR may not always accurately discern the complex anatomical relationships between the transverse processes of the vertebra and the sacrum. This limitation leads to a high rate of misclassification- 35.2% of cases in their study, which can have direct consequences on patient management and outcomes. The misclassifications are attributed to the inability of AP-LPR to detail incomplete joint-like structures or bony union structures and remnants of sclerotic bands, crucial for proper classification under the Castellvi system.
The proposed approach involves a standardized, multidisciplinary-developed classification system that incorporates the use of CT-CRIs. This method promises a detailed view of the osseous anatomical structures necessary for accurate identification and classification of LSTVs. By adopting CT-CRIs as the standard diagnostic tool, the researchers aim to enhance the diagnostic accuracy and reduce the variability and subjectivity associated with AP-LPR.
The implementation of this system included developing comprehensive training materials for radiologists, which comprised schematic diagrams, detailed descriptions, and representative imaging examples. The effectiveness of the new classification system was validated in a study involving one hundred patient examinations, where CT-CRIs were compared with AP-LPR results. The study found significant improvements in diagnostic consistency post-training, with inter-rater agreement levels showing marked enhancement.
Liawrungrueang et al. [6] addressed a critical challenge in medical diagnostics- the reliable detection, classification, and grading of lumbar intervertebral disc degeneration (IDD), a condition linked with chronic low back pain. Traditionally, IDD diagnosis has relied heavily on MRI assessments interpreted by radiologists, a method susceptible to significant variability due to subjective interpretation and potential human error. To combat this, the researchers introduced a sophisticated artificial neural network model, specifically a deep convolutional neural network (CNN), designed to automate and enhance the accuracy of IDD diagnostics through MRI images.

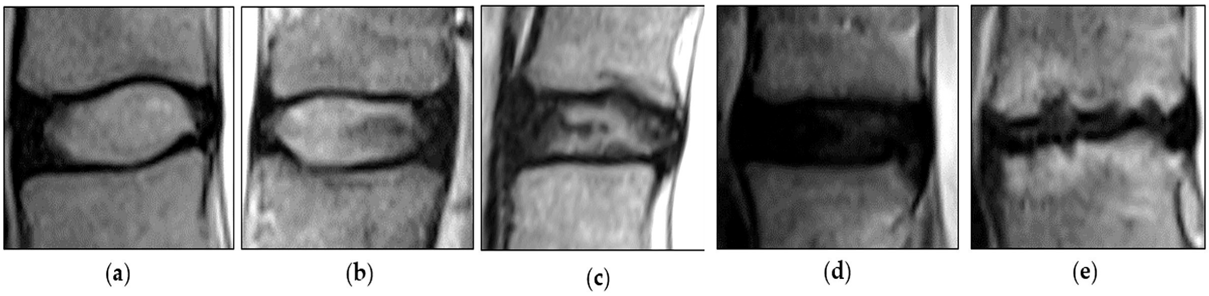


Fig 4. The Pfirrmann grading system: grade I (a), grade II (b), grade III (c), grade IV (d), and grade V (e).[6]

The study's innovation lies in its use of deep learning to provide a standardized, objective, and replicable tool for medical imaging analysis. This method represents a significant departure from traditional diagnostic practices, where the interpretation of MRI scans can vary widely among clinicians, leading to inconsistencies in patient treatment plans. The CNN model developed by the team was meticulously trained on a dataset of 800 annotated MRI images, with a separate set of two hundred images used for validation. This training involved detailed labeling and annotation of the MRI scans by experienced radiologists based on the Pfirrmann grading system, a recognized scale for assessing the severity of IDD (Fig. 4).
The application of the CNN model also offers a way to reduce the workload on radiologists, who often face large volumes of imaging data that need to be meticulously examined. By automating the initial

stages of image assessment, the model allows specialists to focus on more complex cases or on validating the AI's conclusions, thereby enhancing overall diagnostic efficiency.

This work is a major step forward in medical imaging technology, setting the stage for future progress. By improving how AI is used, it could transform diagnostic procedures in many medical areas, including orthopedics. The study shows that AI can speed up the diagnosis process, which is important for quickly identifying conditions like IDD that have a significant impact on patient's health.

3.Background

3.1 Magnetic Resonance Imaging:

Magnetic Resonance Imaging (MRI) is a non-invasive imaging technology that produces three dimensional detailed anatomical images. It is often used for disease detection, diagnosis, and treatment monitoring. MRIs uses powerful magnets which produce a strong magnetic field that forces protons in the body to align with that field. When a radiofrequency current is then pulsed through the patient, the protons are stimulated, and spin out of equilibrium, straining against the pull of the magnetic field. When the radiofrequency field is turned off, the MRI sensors can detect the energy released as the protons realign with the magnetic field. The time it takes for the protons to realign with the magnetic field, as well as the amount of energy released, changes depending on the environment and the chemical nature of the molecules. Physicians can tell the difference between several types of tissues based on these magnetic properties. To obtain an MRI image, a patient is placed inside a large magnet and must remain very still during the imaging process in order not to blur the image. Contrast agents (often containing the element Gadolinium) may be given to a patient through the veins before or during the MRI to increase the speed at which protons realign with the magnetic field. The faster the protons realign, the brighter the image [7].

MRI offers several advantages over other imaging techniques such as computed tomography (CT) for spinal imaging. On MRI, Soft tissue contrast is better, which allows parts of the disc to be distinguished from one another. in addition, MRI visualizes the ligaments better. MRI also offers better visualization of the vertebral marrow and contents of the spinal canal.

A disadvantage of MRI is that it cannot directly visualize cortical bone, which does not have mobile protons and produces a black “signal void” on magnetic resonance images. When bony anatomy is critical, CT is preferable. In patients who have had acute trauma, for example, CT may better depict fractures, especially of the posterior elements [14].

MRI is commonly used for diagnosing and investigating spine conditions. It helps doctors classify and grade the condition, showing its severity and guiding treatment decisions, whether surgery or non-surgical methods are needed. The Pfirrmann grading system is often used by radiologists and orthopedic specialists to assess these conditions [6]. However, that process is time-consuming for radiologists who are already overloaded with work related to prediction and having to classify the level of each patient. Moreover, 70% of all medical errors in diagnostic radiography are “missed findings” according to an examination of those errors [18]. This significant error rate demonstrates how difficult the "detection process” is for humans.

3.2 Convolutional Neural Networks-CNN:

Convolutional Neural Networks (CNNs) are a sophisticated subset of deep learning frameworks specifically engineered to analyze visual data efficiently. CNNs are adept at automatically learning spatial hierarchies of features through multiple layers, making them particularly effective for tasks involving image and video processing. The architecture of a CNN usually involves several convolutional layers which help in extracting various features from the images. These layers are often mixed with pooling layers that reduce the spatial size of the representation, thus reducing the parameter counts and computation in the network. This setup is typically followed by one or more fully connected layers that perform the high-level reasoning in the neural network (Fig .5).

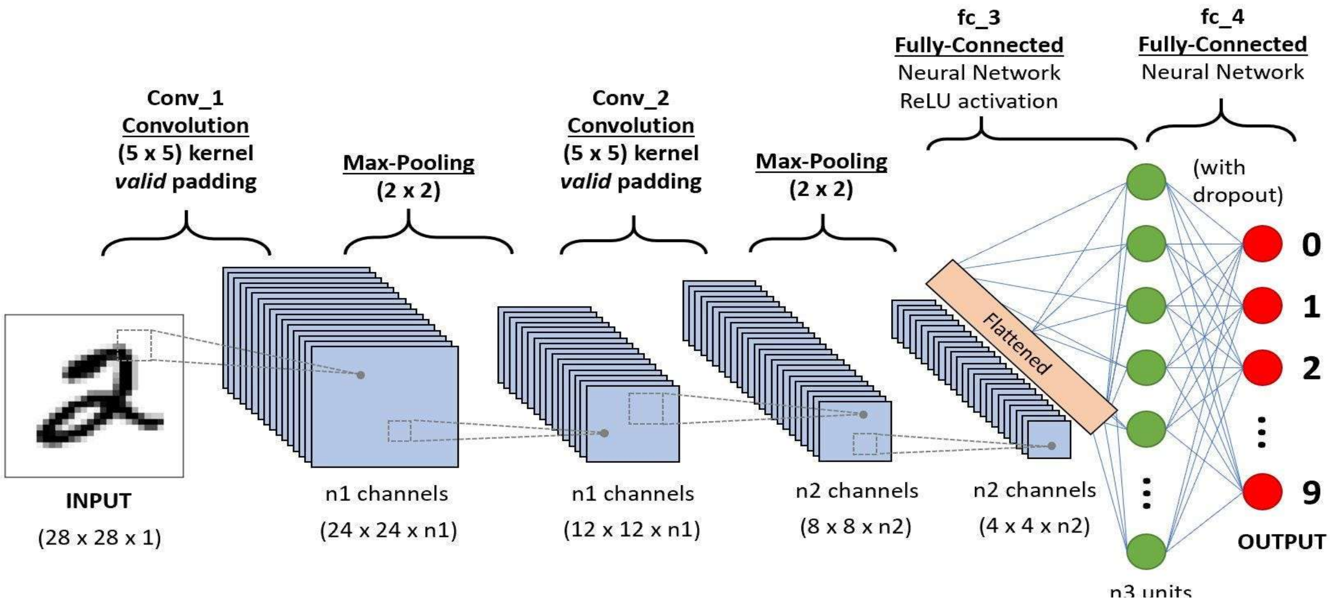


Fig 5 A traditional CNN sequence to classify handwritten digits

Convolutional layers: responsible for extracting spatial hierarchies of features from input images. They use filters or kernels (small matrices) that slide across the image. Each filter detects a specific feature like edges, curves, corners, or textures. Each filter result in a feature map which is a new image that highlights where did the filter detect certain features in the original image. CNNs usually use many filters in a single convolutional layer, each detecting a different feature we want to detect, which leads to multiple feature maps giving the network a better understanding of the input image. By stacking multiple convolutional layers, the network can detect more complex features, starting with simple patterns in the early layers and building up to more complex shapes and objects in deeper layers.

Pooling layers: max pooling works by taking a small region from the feature map (usually 2X2) and selecting the maximum value from this region. This is done for every region as it slides over the entire feature map. Max pooling plays a crucial role in reducing the spatial dimensions of the feature maps while retaining essential information.

Fully connected layers: the final layers of a CNN. They take the high-level features learned from the previous layers and map them to specific classes or outputs. By leveraging these extracted features, fully connected layers can make predictions or perform classification tasks by assigning weights to each feature they received and determine which class is best suited as a result.[12]

Additionally, Neural networks leverage several types of activation functions to introduce non-linearities and enable learning complex patterns. Non-linearity is crucial for modeling real-world relationships between objects that do not have straightforward, linear connections and uses an activation function that introduces nonlinearity.
The most widely used activation function in CNNs, is ReLU as defined in Eq. 1.

$$f(x) = \max(0, x)$$

The ReLU function thresholds inputs at zero, outputting 0 for negative values and passing positive values unchanged, acting linearly with a gradient of 1 for inputs above zero. This characteristic is essential in avoiding the vanishing gradient problem during backpropagation. [13]

3.3 Transfer Learning

Transfer learning (TL) with convolutional neural networks aims to improve performances on a new task by leveraging the knowledge of similar tasks learned in advance. It has made a major contribution to medical image analysis as it overcomes the data scarcity problem as well as it saves time and hardware resources. (fig .6).
in TL with CNN for medical image classification, a medical image classification (target task) can be learned by leveraging the generic features learned from the natural image classification (source task) where labels are available in both domains[11].to put simply, we can use the CNN that was trained on ImageNet for our advantage and use it for our classification of lumbar spine degenerations.

Inductive Transfer Learning is used when the tasks between the source and target domains are different, though they may share some similarities. For example, a neural network trained on a large image dataset like ImageNet may be adapted to perform a specific task such as medical image diagnosis or object detection in different settings.

In multitask learning, a single model is trained to tackle multiple tasks simultaneously, such as recognizing objects and classifying scenes in images. This approach leverages shared representations across tasks, allowing the model to generalize better and learn more efficiently by capturing common features and patterns.

Practical Considerations

When applying transfer learning, the model is typically fine-tuned on the new dataset to optimize its performance for the specific task. This involves adjusting hyperparameters such as the learning rate and the number of training epochs, as the model's initial weights have already been learned from a different task. Fine-tuning allows the model to adapt its knowledge to the new task while retaining valuable features learned during the initial training.

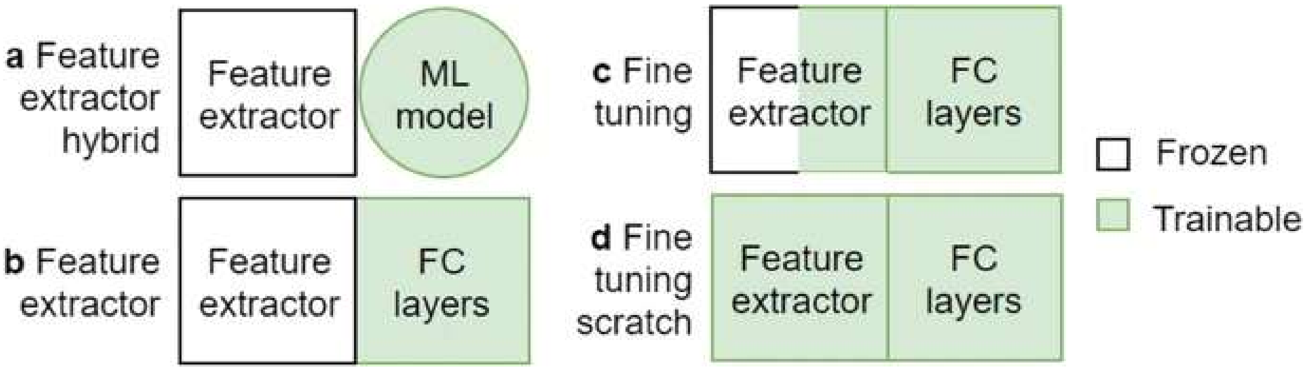


Fig 6. Four types of transfer learning approach. The last classifier block needs to be replaced by a thinner layer or trained from scratch (ML: Machine learning; FC: Fully connected layers) [11]

3.4 Loss function

in machine learning, loss function measures how well a model's predictions match the actual data. It quantifies the difference between the predicted values and the actual values - the true probability value of each sample will be 1 for the class that it belongs to and 0 for the rest of the classes. Loss function provides a metric that the model aims to minimize during training. The loss function guides the optimization process by indicating how far off the predictions are, the goal is to adjust the model parameters to reduce this loss.

Cross Entropy loss function

Cross-entropy, also known as logarithmic loss or log loss, is a loss function used in machine learning to measure the performance of a classification model.
cross-entropy measures the difference between the discovered probability distribution of a classification model and the predicted values.
The cross-entropy loss function is used to find the optimal solution by adjusting the weights of a machine learning model during training. The objective is to minimize the error between the actual and predicted outcomes. A lower cross-entropy value indicates better performance.

For multi-class classification problems, where an instance could belong to one of many classes, the cross-entropy loss is generalized to Eq. 2:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} * \log(\hat{y}_{i,c})$$

Eq.2

where:

- C is the number of classes.
- $y_{i,c}$ is a binary indicator (0 or 1) if class label c is the correct classification for instance i.
- $\hat{y}_{i,c}$ is the predicted probability of instance i being of class c.
- N is the number of instances in the data set

Each inner sum computes the loss for each class label per sample by taking the logarithm of the predicted probability for the true class and summing these values across all classes. The result is then averaged over all samples [17]

4.Process

4.1 Dataset

in this project we will utilize the dataset of The Radiological Society of North America (RSNA) 2024 Lumbar Spine Degenerative Classification for both training and testing [10]. In collaboration with the American Society of Neuroradiology (ASNR), the Radiological Society of North America (RSNA) has curated a comprehensive dataset for the purpose of advancing the detection and classification of degenerative spine conditions through artificial intelligence (AI). The dataset, made available via the Kaggle competition "RSNA 2024 Lumbar Spine Degenerative Classification," provides a rich resource for exploring the potential of AI in medical imaging, particularly in the domain of lumbar spine MR images. The dataset comprises a collection of MRI images annotated with severity scores for five specific degenerative lumbar spine conditions: Left Neural Foraminal Narrowing, Right Neural Foraminal Narrowing, Left Subarticular Stenosis, Right Subarticular Stenosis, and Spinal Canal Stenosis. Each of these conditions has been evaluated at the intervertebral disc levels L1/L2, L2/L3, L3/L4, L4/L5, and L5/S1.

The severity scores provided in the dataset categorize the conditions into three levels: Normal/Mild, Moderate, or Severe, allowing for nuanced analysis and classification. To establish the ground truth dataset, the RSNA challenge planning task force gathered imaging data from nine sites across five continents. This multi-institutional, expertly curated dataset aims to improve the standardized classification of degenerative lumbar spine conditions and facilitate the development of tools for accurate and rapid disease classification. All the imaging datasets have been manually labeled by more than fifty experts worldwide whose names are motioned in the official Kaggle project website [10]

The dataset contains 147,320 files in the size of 35.34GB, and contains both CSV files for results and classification, and DCM files of the MRI images**.** The data is divided into test images, train images, and train.csv for results.
To display the DICOM files, we wrote a Python script to convert the files to PNGs which will be used in our project. Examples are seen in Figure 7.

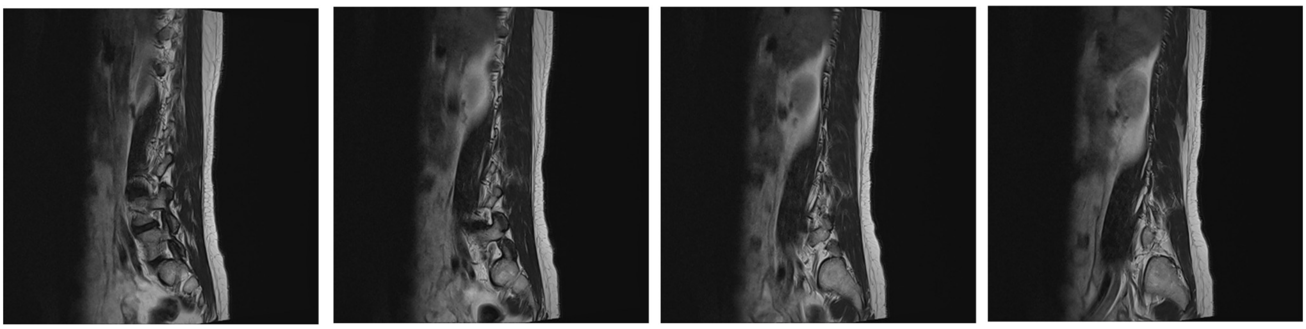


Fig 7 MRI of the lower-back from Kaggle dataset.[9]

4.2 Hyper-Parameters Optimization

Classification is a supervised machine learning method where the model tries to predict the correct label or class of a given input data. In classification, the model is fully trained using the training data, where images are labeled correctly, and then it is evaluated on test data before being used to perform prediction on new unseen data.[15]

Based on sensitivity, specificity, precision, and recall, we will evaluate the performance of the classification system .

The following measures were evaluated:

- **true positive (tp)** = The model will predict lumbar spine degeneration and the lumbar spine MRI will be classified as lumbar spine degeneration.

- **true negative(tn)** = The model will predict no lumbar spine degeneration, and the lumbar spine MRI will be classified as no lumbar spine degeneration.

- **false positive(fp)** = The model will predict lumbar spine degenerations, but the lumbar spine MRI will be classified as no lumbar spine degenerations.

- **false negative(fn)** = The model will predict no lumbar spine degeneration, but the lumbar spine MRI will be classified as lumbar spine degeneration.

Secondly, we will count each category instances and measure the following parameters [15]:

- **Accuracy –** The percentage of "true" predictions out of all the predictions = $\frac{(tp + tn)}{(tp + tn + fp + fp)}$

- **Precision –** Out of all the positive predictions, how many of them were correct = $\frac{(tp)}{(tp + fp)}$

- **Recall /sensitivity** -Determines how good was the model at predicting real "yes" events, which means that we count the true positive instances out of all the instances where the MRI will be classified as a **lumbar spine degeneration** = $\frac{(tp)}{(tp + fn)}$

-**Recall/specificity** - Determines how good was the model at predicting real "no" events, which means that we count the true negative instances out of all the instances where the MRI will be classified as a **no lumbar spine degeneration** =$\frac{(tn)}{(tn + fp)}$

- **F1-score** - Used for imbalanced datasets (where the number of instances of each class/label is approximately the same).F1 score is the harmonic mean of precision and recall = $\frac{(2 * Recall * Precision)}{(Recall + Precision)}$

One of the challenges that classification methods need to consider is overfitting the data. Overfitting data means that the model gives accurate predictions for training data but not for new data [16].

Overfitting occurs when the model cannot generalize and fits too closely to the training dataset instead. Overfitting happens due to several reasons, for example: The training data size is too small and does not contain enough data samples to accurately represent all possible input data values, The training data contains large amounts of irrelevant information (called noisy data), the model trains for too long on a single sample set of data, The model complexity is high so it learns the noise within the training data[16].

4.3 DenseNet

DenseNet, short for Densely Connected Convolutional Network, is a CNN architecture designed to address several limitations of previous deep learning models, particularly those encountered in very deep networks like ResNet. Developed by Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger.[8]

DenseNet enhances traditional network designs by connecting each layer directly to every other layer in a feed-forward manner (Fig. 8), where the output of each layer is fed into all following layers. This connectivity pattern offers several advantages: It boosts the efficiency of the network by promoting feature reuse and eliminating the redundancy of relearning features. As a result, reducing the number of parameters, making the network more compact and reduce information loss.
Unlike ResNet, which merges outputs from skip connections using addition, DenseNet concatenates these outputs, preserving and utilizing all incoming information (Fig.9).

Moreover, the direct connections ensure that all layers have access to the gradients from the loss function, simplifying the training process.

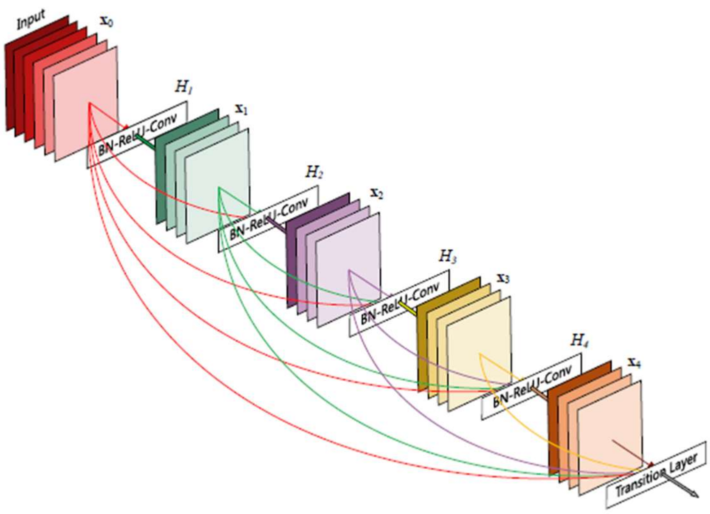


Fig 8 5-layer dense block with a growth rate of k = 4. Each layer takes all preceding feature-maps as input.[8]

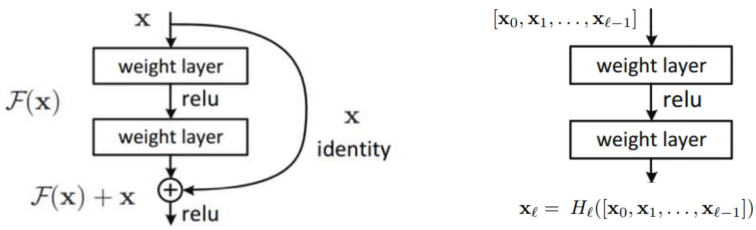


Fig 9 on the left: ResNet layers, each layer (F(x)+x) uses the identity of the previous layer(X)
On the right: DenseNet layer, each layer uses the concatenation of all the previous layers [8]

DenseNet has been evaluated on various highly competitive object recognition benchmarks, such as ImageNet. The results showed significant improvements over other models like ResNet. DenseNet showed lower validation errors and test errors compared to ResNets variations [8] There are several types of DenseNet according to the layers we are using. We will be using DenseNet-121 (Table 1).

Layers	Output Size	DenseNet-121
Convolution	112×112	
Pooling	56×56	
Dense Block (1)	56×56	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$
Transition Layer (1)	56×56	
	28×28	
Dense Block (2)	28×28	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$
Transition Layer (2)	28×28	
	14×14	
Dense Block (3)	14×14	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 24$
Transition Layer (3)	14×14	
	7×7	
Dense Block (4)	7×7	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 16$
Classification Layer	1×1	

Table 1 DenseNet architecture [8]

As we see, the DenseNet architecture consists of few layers: convolution layer, pooling layer, dense block layers, transition layers and classification layer.

Convolution layer and pooling layer has already been described in the CNN section of the project.

The first convolution layer uses a 7X7 convolution with stride of 2. the first pooling convolution uses 3X3 max pool with stride of 2.

Dense block layer - consists of 6 layers of 1X1 and 3X3 convolutions.in DenseNet-121 we will use 6 layers in the first dense block, 12 layers in the second dense block,24 layers in the third dense block and 16 layers in the last block.

Transition layers - use 1X1 convolution followed by 2X2 average pooling between two adjacent dense blocks.

Classification layer - At the end of the last dense block, a 7X7 global average pooling is performed and then a softmax classifier is attached**.** Uses 1000D fully connected which means that there are 1000 neurons at the end which are all connected.[8]

One of the main problems in ResNet that DenseNet solves is the Vanishing Gradient problem, The vanishing gradient problem is a common issue in deep neural networks, where gradients used for training diminish as they are backpropagated through the layers. This can lead to slow convergence or even failure to train effectively. The problem occurs because the gradient of the loss function with respect to weights becomes increasingly small.

DenseNet's architecture ensures that gradients can flow directly from the loss function to any previous layer. By concatenating feature maps from all preceding layers, DenseNets not only improve the flow of gradients but also enhance feature reuse and propagation. This architecture enabling the model to train deeper networks more efficiently without suffering from the degradation of information.

4.4 Hyper-Parameters

Hyperparameters in CNNs are settings selected prior to training that determines how the model learns. Unlike weights and biases, hyperparameters are not derived directly from the training data but are set to enhance the model's performance.

For our project, we'll use the following configurations:

1. **Learning rate**: Ranges from $1 * 10^{-3}$ to $1 * 10^{-7}$ This determines the size of the steps taken during gradient descent to minimize the loss function. We'll adopt a variable learning rate approach, testing different values to see which minimizes error most effectively based on our results graphs (loss and accuracy graphs).

2. **Batch size**: This is the number of training samples used before the model updates its parameters. We will experiment with batch sizes of 32 and 64, balancing between computational efficiency and the stability of the learning updates.

3. **Epochs**: This refers to the number of complete passes through the training dataset. To fine-tune our model, we will compare the effects of running 50 and 100 epochs, observing the trade-off between learning performance and the risk of overfitting.

4. **Dropout rate**: As a regularization method, dropout randomly disables a percentage of neurons during training to prevent the model from relying too heavily on any small set of neurons. We will test dropout rates of 0.2 and 0.5 to identify the best option for mitigating overfitting.

These hyperparameters will guide the training of our CNN as we aim to classify different types of Lumbar Spine Degeneration effectively.

5. Expected achievements

In this project we expected to build a system that will be able to classify MRI scans into five specific degenerative lumbar spine conditions: Left Neural Foraminal Narrowing, Right Neural Foraminal Narrowing, Left Subarticular Stenosis, Right Subarticular Stenosis, and Spinal Canal Stenosis. Each of these conditions will be evaluated at the intervertebral disc levels L1/L2, L2/L3, L3/L4, L4/L5, and L5/S1.

One of the challenges in diagnosing lumbar spine degenerations from MRI scans is the variability in how different doctors interpret the images. Each doctor may approach the problem from a slightly different perspective, and there is not a universally accepted "gold standard" for diagnosis, leading to inconsistencies.
Our goal is to develop a reliable and accurate system that simplifies this process as much as possible. We aim for the system to achieve at least 80% accuracy, ensuring it is both user-friendly and efficient. Additionally, we hope our system will reduce the time doctors spend analyzing MRI scans, helping to speed up diagnosis so patients can receive their results faster than the current methods.

The objectives of this project will be accomplished by training a CNN to accurately identify and differentiate various stages of lumbar spine degeneration. The CNN will be trained using the Kaggle's dataset which consists of MRI scans annotated with severity levels (normal/mild, moderate, severe). To enhance the learning, we will implement the DenseNet architecture which enhances traditional network designs by connecting each layer directly to every other layer in a feed-forward manner. Additionally, we will leverage transfer learning to accelerate the training process by utilizing features pre-learned from natural image classification (such as the ImageNet dataset), allowing the model to converge faster and perform better with fewer data.

For our project to be successful:

- build a working network based on DenseNet.

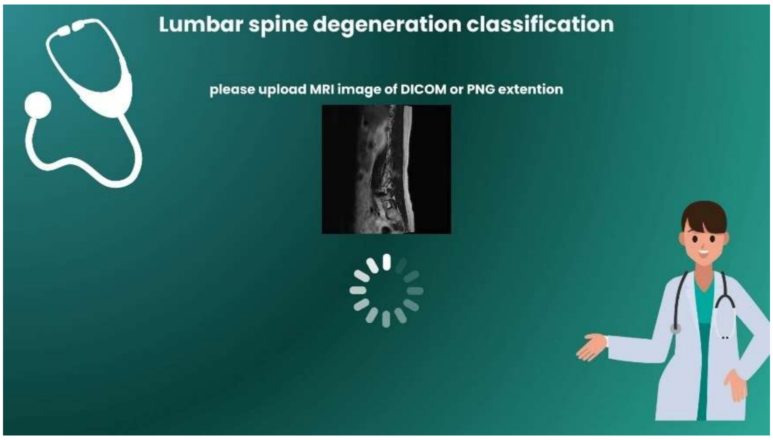
- achieve an accuracy better than 0.8.

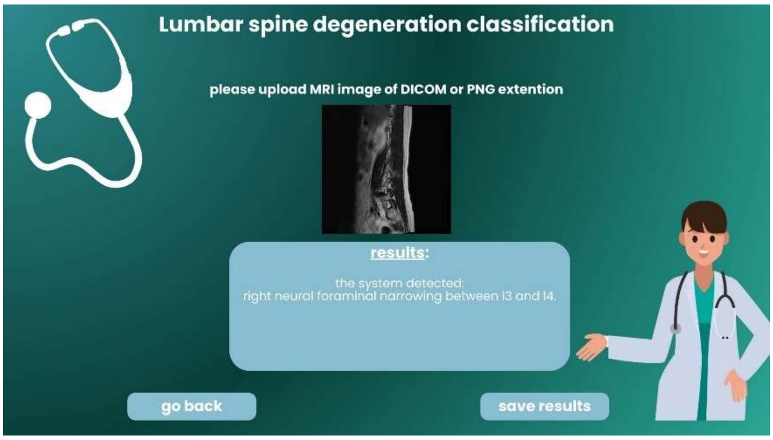
- to grade each lumbar spine degeneration in the scan.

5.1 GUI


Lumbar spine degeneration classification
please upload MRI image of DICOM or PNG extention
Upload


Lumbar spine degeneration classification
please upload MRI image of DICOM or PNG extention
Upload


Lumbar spine degeneration classification
please upload MRI image of DICOM or PNG extention


Lumbar spine degeneration classification
please upload MRI image of DICOM or PNG extention
results:
the system detected:
right neural foraminal narrowing between l3 and l4.
go back
save results

5.3 system testing

Case number	Test case	Expected result
1	Insert wrong image format	Error message: "Wrong format"
2	Insert DICOM or PNG format	The system will allow to press the "upload" button.
3	Insert expected picture format with no lumbar spine degeneration	The system will allow to press the "upload" button and prompting "no degeneration" in the result screen.
4	Press the "upload" button without an image uploaded	No action since the "upload" button will not be available to press
5	Upload viable image, with lumbar spine degeneration	The system will allow to press "upload" and show the classification result
6	Press "back" button on the last GUI screen	The system will go to the first GUI screen
7	Press "save result" button on the last GUI screen	The system will prompt local save option

6. Conclusions

In the first phase of this project, we built the necessary background knowledge to fully understand the problem and explore potential solutions. The first step was to read articles explaining how lumbar spine degeneration is diagnosed and the challenges it faces. We found that one of the biggest issues is the lack of a "gold standard" for diagnosis, which leads to variability in how different doctors analyze the same MRI scans, potentially delaying the diagnosis process.

First, we studied the background of lumbar spine degeneration to understand the problem fully. We learned about how MRI scans work and how they contribute to diagnosing these conditions. Once we had a solid understanding of the issue, we explored viable solutions, with a focus on using artificial intelligence (AI) in medicine. Specifically, we researched how machine learning could be applied to classify lumbar spine degenerations from MRI scans.

Next, we explored convolutional neural networks (CNNs) and learned how they can be used to solve our problem by training on labeled MRI scans in the gives Kaggle's dataset and applying that knowledge to classify new, unlabeled scans. We also studied the DenseNet architecture, which improves deep learning models and offers advantages over other architectures like ResNet. Finally, we learned about transfer learning and how it can speed up the training process by using general features learned from natural image classification on the ImageNet dataset.

References

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

2. Miskin, N., Gaviola, G. C., Huang, R. Y., Kim, C. J., & Thomas. (2022). *Standardized classification of lumbar spine degeneration on magnetic resonance imaging reduces intra- and inter-subspecialty variability*. Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA.

3. National Institute of Arthritis and Musculoskeletal and Skin Diseases. (2023, February). *Back pain*. https://www.niams.nih.gov/health-topics/back-pain
4. Gille, O., Bouloussa, H., Mazas, S., Vergari, C., Challier, V., Vital, J.-M., Coudert, P., & Ghailane, S. (2017). *A new classification system for degenerative spondylolisthesis of the lumbar spine*. *European Spine Journal*, 26(10), 2521-2527.
5. Hou, L., Bai, X., Li, H., Gao, T., Li, W., Wen, T., He, Q., Ruan, D., Shi, L., & Bing, W. (2020). *Lumbar plain radiograph is not reliable to identify lumbosacral transitional vertebra types according to Castellvi classification principle*. *BMC Musculoskeletal Disorders, 21*(333). https://doi.org/10.1186/s12891-020-03358-3
6. Liawrungrueang, W., Kim, P., Kotheeranurak, V., Jitpakdee, K., & Sarasombath, P. (2023). *Automatic detection, classification, and grading of lumbar intervertebral disc degeneration using an artificial neural network model*. https://www.mdpi.com/2075-4418/13/4/663

7. National Institute of Biomedical Imaging and Bioengineering. (n.d.). Magnetic Resonance Imaging (MRI). National Institutes of Health. https://www.nibib.nih.gov/science-education/science-topics/magnetic-resonance-imaging-mri

8. Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2018). Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, 4700-4708. https://doi.org/10.1109/CVPR.2017.243

9. Kaggle. (2024). *RSNA 2024 lumbar spine degenerative classification*. Kaggle. Retrieved from https://www.kaggle.com/competitions/rsna-2024-lumbar-spine-degenerative-classification

10. Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234-241. Springer, Cham, 2015.

11. Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., & Ganslandt, T. (2022). Transfer learning for medical image classification: A literature review. BMC Medical Imaging, 22(69). https://doi.org/10.1186/s12880-022-00793-7

12. AI and Insights. (2021, June 10). Convolutional neural networks (CNNs) in computer vision. *Medium*. https://medium.com/@AIandInsights/convolutional-neural-networks-cnns-in-computer-vision-10573d0f5b00

13. DataCamp. (n.d.). *Introduction to activation functions in neural networks*. DataCamp. https://www.datacamp.com/tutorial/introduction-to-activation-functions-in-neural-networks

14. Jarvik, J. G., & Deyo, R. A. (2002). Diagnostic evaluation of low back pain with emphasis on imaging. *Annals of internal medicine*, *137*(7), 586–597. https://pubmed.ncbi.nlm.nih.gov/12353946/

15. DataCamp. (2024, August). *Classification in machine learning: A guide for beginners*. DataCamp. https://www.datacamp.com/blog/classification-machine-learning

16. Amazon Web Services. (2024). *What is overfitting?* Amazon Web Services. https://aws.amazon.com/what-is/overfitting/

17. GeeksforGeeks. (2024, January 3). *What is Cross-Entropy Loss Function?* https://www.geeksforgeeks.org/what-is-cross-entropy-loss-function/

18. Carrino, J. A., Morrison, W. B., Zou, K. H., Steffen, R. T., Manoni, G. R., & Schweitzer, M. E. (2001). Lumbar disc herniation: Accuracy of MR imaging findings. American Journal of Roentgenology, 177(5), 1205-1212. https://doi.org/10.2214/AJR.13.11493

19. Project GitHub Repository: https://github.com/danielZada97/Capstone-Project-24-2-R-1