
FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

**Mustafa Al-Rubaye**

# DEEP LEARNING-BASED LOWER BACK PAIN CLASSIFICATION AND DETECTION FROM T2-WEIGHTED MAGNETIC RESONANCE IMAGES

Master's Thesis
Degree Programme in Biomedical Engineering
January 2023

**Al-Rubaye M. (2023) Deep Learning-Based Lower Back Pain Classification and Detection From T2-Weighted Magnetic Resonance Images.** University of Oulu, Degree Programme in Biomedical Engineering, 52 p.

# ABSTRACT

**Lower back pain (LBP) is a common physiological condition that affects 50-80% of the adult population at some point in their lives. For example, the economic load of LBP in Sweden was estimated to be approx. at €740 million in 2011.**

**In LBP diagnostics, magnetic resonance imaging (MRI) is often used. MRI is used to visualize the structures in the lumbar region of the spine such as disks, bones, and spaces between the vertebral bones where nerves pass through. The lumbar spine refers to the lowest five vertebrae and intervertebral discs of the spine. MRI provides a detailed picture of the lumbar spine to get visual confirmation of any abnormalities potentially related to LBP to support the diagnosis process.**

**The goal of this thesis was to investigate visual patterns related to LBP in $T_2$-weighted MR images measured with a fast spin-echo sequence on a GE Healthcare Signa HDxt 1.5 T MRI system. A convolutional neural network was used to classify MRIs into symptomatic and asymptomatic cases and to develop a fully automated pain prediction process. A total of 526 MRI examinations with supporting pain questionnaires from the Northern Finland Birth Cohort 1966 (NFBC1966) were used. Three different datasets were created for the experiments: i) a dataset with mid-sagittal slices from the center of the spine from each examination, ii) a dataset with mid-sagittal slices and its immediate neighboring slices, and similarly, iii) a dataset with five middle-most sagittal slices. In each dataset, individual slices were considered as independent samples, i.e., inputs for the classification method.**

**The developed classification method yielded the best results when the input dataset comprised of three middle-most slices (Balanced Accuracy score (BACC) of 0.709 $\pm$ 0.011, Average Precision (AP) of 0.467 $\pm$ 0.025, and Area Under Receiver Operating Characteristic curve (ROC-AUC) of 0.740 $\pm$ 0.008). The baseline model trained using only the mid-sagittal slice for classification yielded the lowest classification scores (BACC of 0.546 $\pm$ 0.032, AP of 0.403 $\pm$ 0.007, and ROC-AUC of 0.667 $\pm$ 0.008) followed by the model trained with the dataset with five middle-most slices (BACC of 0.675 $\pm$ 0.008, AP of 0.369 $\pm$ 0.009, and ROC-AUC of 0.619 $\pm$ 0.011).**

**To conclude, this work suggests that the developed deep learning-based classification pipeline could be used for LBP diagnostics of lumbar spine MRI. LBP diagnostics is heavily based on degenerative MRI findings and deep learning has the potential to supplement these visual assessments objectively. The developed method could be helpful, for example, in identifying negative cases in order to enhance the workflow of routine diagnostic imaging tasks.**

**Keywords: deep learning, image classification, magnetic resonance imaging, lumbar spine, low back pain**

**Al-Rubaye M. (2023) Alaselkäkivun luokittelu ja havainnointi T2-painotetuista magneettikuvista syväoppimista hyödyntäen.** Oulun yliopisto, Lääketieteen tekniikan maisteriohjelma, 52 s.

# TIIVISTELMÄ

**Alaselkäkipu on yleinen fysiologinen tila, joka vaikuttaa 50:stä 80:een %:iin aikuisväestöstä jossain vaiheessa heidän elämäänsä. Ruotsissa alaselkäkipuun liittyvän taloudellisen kuormituksen on arvioitu olleen noin 740 miljoonaa euroa vuonna 2011.**

**Alaselkäkivun syyn etsimiseen käytetään tyypillisesti magneettikuvausta (MRI). MRI:tä käytetään lannerangan alueen rakenteiden, kuten levyjen, luiden ja selkärangan luiden välisten tilojen, joissa hermot kulkevat, visualisoimiseen. Lannerangalla tarkoitetaan selkärangan viittä alinta nikamaa ja levyä. MRI tarjoaa diagnoosin tukemiseksi yksityiskohtaisen kuvan lannerangasta mahdollistaen alaselkäkipuun mahdollisesti liittyvien poikkeamien visuaalisen tarkastelun.**

**Tämän opinnäytetyön tavoitteena oli tutkia alaselkäkipuun liityviä muutoksia $T_2$-painotetuissa magneettikuvissa, jotka kuvattiin GE Healthcare Signa HDxt 1,5 T magneettikuvauslaitteistolla nopeaa spin-kaikusekvenssiä käyttäen. Kuvien luokitteluun käytettiin konvoluutioneuroverkkoja oireellisiin ja oireettomiin tapauksiin täysautomatisen kivun ennustusmenetelmän kehittämiseksi. Aineistona käytettiin yhteensä 526 tutkimusta Pohjois-Suomen syntymäkohortista 1966 (NFBC1966). Testejä varten luotiin kolme erilaista aineistoa: i) keskisagittaaliset viipalekuvat, ii) keskisagittaaliset viipalekuvat ja niiden naapuriviipaleet, sekä vastaavasti iii) viisi keskimmäisintä viipalekuvaa, joita hyödynnettiin itsenäisinä näytteinä, eli luokitusmenetelmän syötteinä.**

**Kehitetty luokitusmenetelmä tuotti parhaat tulokset kun syötejoukkona olivat keskisagittaaliset viipalekuvat ja niiden naapuriviipaleet (Balanced Accuracy score (BACC) 0,709 $\pm$ 0,011, Average Precision (AP) 0,467 $\pm$ 0,025, ja Area Under Receiver Operating Characteristic curve (ROC-AUC) 0,740 $\pm$ 0,008). Keskisagittaalisten viipalekuvien avulla koulutettu vertailumalli tuotti alhaisimmat luokittelutulokset (BACC 0.546 $\pm$ 0.032, AP 0.403 $\pm$ 0.007, and ROC-AUC 0.667 $\pm$ 0.008), ja seuraavaksi paras malli oli viidellä keskimmäisellä viipalekuvalla koulutettu malli (BACC 0.675 $\pm$ 0.008, AP 0.369 $\pm$ 0.009, and ROC-AUC 0.619 $\pm$ 0.011).**

**Tämä työ antaa viitteitä siitä, että syväoppimiseen perustuvaa menatelmää voittaisiin käyttää lannerangan MRI-aineistosta suoritettavaan alaselkäkivun diagnosointiin. Alaselkäkivun diagnostiikka perustuu vahvasti MRI-rappeumalöydöksiin, ja syväoppimisella on edellytyksiä täydentää objektiivisella tavalla näitä visuaalisia arvioita. Kehitetystä menetelmästä voisi olla apua esimerkiksi negatiivisten tapausten tunnistamisessa rutiininomaisten diagnostisten kuvantamistehtävien työnkulun tehostamiseksi.**

**Avainsanat: syväoppiminen, kuvan luokittelu, magneettikuvaus, lanneranka, alaselkäkipu**

# TABLE OF CONTENTS

ABSTRACT
TIIVISTELMÄ
TABLE OF CONTENTS
FOREWORD
LIST OF ABBREVIATIONS AND SYMBOLS
1. INTRODUCTION ..... 8
2. BACKGROUND ..... 9
2.1. Lumbar Spine ..... 9
2.2. Low Back Pain ..... 10
2.2.1. Economical Load ..... 11
2.2.2. Diagnosis ..... 11
2.2.3. Management and Treatment ..... 13
2.3. Magnetic Resonance Imaging with $T_2$-Weighted Contrast ..... 14
2.4. Deep Learning ..... 17
2.4.1. Deep Learning in Medical Image Analysis ..... 18
2.4.2. Artificial Neural Networks ..... 18
2.4.3. Convolutional Neural Networks and Residual Neural Networks ..... 19
2.4.4. Supervised and Unsupervised Learning ..... 20
2.4.5. Optimization ..... 21
2.4.6. Overfitting and Cross-Validation ..... 22
2.4.7. Regularization Methods ..... 23
2.4.8. Transfer Learning ..... 24
2.5. Image Segmentation ..... 24
2.6. Image Classification ..... 25
2.7. Evaluation Metrics ..... 25
3. MATERIALS AND METHODS ..... 27
3.1. Cohort Data ..... 27
3.2. Data Annotation ..... 28
3.3. Segmentation of Cropped Images ..... 30
3.3.1. Data Processing and Augmentation ..... 30
3.3.2. Model Training and Validation Scheme ..... 30
3.4. Classification: Experimental Setup ..... 30
3.4.1. Data Splitting ..... 30
3.4.2. Experiments ..... 31
3.5. Deep Learning Framework and Hyperparameters ..... 33
4. RESULTS ..... 35
4.1. Segmentation of Lumbar Region ..... 35
4.2. Effect of Dataset Size in Low Back Pain Detection ..... 35
5. DISCUSSION ..... 40
5.1. Main Findings ..... 40
5.2. Related Work ..... 41
5.3. Limitations and Future Work ..... 42
6. SUMMARY ..... 43
7. REFERENCES ..... 44

# FOREWORD

For humans, learning just comes naturally starting at birth and the process of learning continues until death. The learning process as a consequence happens all around from the interactions with people and the surroundings. Machine learning is a branch of artificial intelligence that studies algorithms that can mimic the same learning process as humans. Machine learning usually uses data to learn using some probabilistic theories. The data used to train machine learning models affects the performance, as typically the trained model cannot outperform the training samples. Overall, human performance is still better than most machine learning models, but this might not be the case for long. An effort has been made recently to develop machine learning algorithms that surpass human performance. In this thesis, the goal was to study whether artificial neural networks can help classify MR images into symptomatic and asymptomatic cases. These types of algorithms allow humans to be more efficient and focus on more complicated tasks.

The aim of this thesis is to investigate if a CNN model could be used to detect lower back pain from $T_2$-weighted MRI of the lumbar spine. Different metrics were used to compare the results (predictions) obtained from the different models with the actual labels.

The work presented in this thesis was conducted at the Research Unit of Health Sciences and Technology at the University of Oulu. The research on the presented topic started on August 2020 and the final results were acquired in the summer of 2022. The research unit and the University of Oulu are acknowledged for providing the infrastructure and the funding for this project.

Dr. Juuso Ketola, Dr. Satu Inkinen, and Mr. Antti Isosalo are acknowledged for supervising this study. Dr. Jaro Karppinen is acknowledged for providing guidance related to lower back pain. Dr. Miika Nieminen is acknowledged for providing fruitful feedback related to the project. I would like to acknowledge the University of Oulu for providing the Northern Finland Birth Cohort 1966 dataset.

Oulu, January 24th, 2023

Mustafa Al-Rubaye

# LIST OF ABBREVIATIONS AND SYMBOLS

| | |
|---|---|
| Adam | Adaptive moment estimation |
| AI | Artificial intelligence |
| ANN | Artificial neural network |
| AP | Average Precision |
| BACC | Balanced accuracy |
| BN | Batch normalization |
| CLBP | Chronic low back pain |
| CNN | Convolutional Neural Network |
| CT | Computed tomography |
| CVAT | Computer Vision Annotation Tool |
| DL | Deep learning |
| DSC | Dice similarity coefficient |
| ETL | Echo train length |
| FID | Free induction decay |
| FOV | Field of view |
| FN | False negatives |
| FP | False positives |
| FPR | False positive rate |
| GPU | Graphics processing unit |
| Grad-CAM | Gradient-weighted Class Activation Mapping |
| IVD | Intervertebral disc |
| KNN | K-Nearest Neighbors |
| LBP | Lower back pain |
| MCC | Matthews correlation coefficient |
| MLP | Multilayer Perceptron |
| MRI | Magnetic resonance imaging |
| NFBC1966 | Northern Finland birth cohort 1966 |
| NMR | Nuclear magnetic resonance |
| NPV | Negative predictive value |
| OOF | Out-of-fold |
| ReLU | Rectified Linear Unit |
| ResNet | Residual Neural Network |
| RF | Radiofrequency |
| ROC | Receiver Operating Characteristic |
| ROC-AUC | Area Under Receiver Operating Characteristic curve |
| SE | Spin echo |
| SMOTE | Synthetic minority oversampling technique |
| SVM | Support Vector Machines |
| T1 | Longitudinal relaxation time |
| T2 | Transverse relaxation time |
| TE | Time of echo |
| TN | True negatives |
| TNR | Specificity score |
| TP | True positives |

| | |
|---|---|
| TPR | True positive rate |
| TR | Repetition time |
| | |
| $\nu$ | Resonance frequency |
| $\gamma$ | Gyromagnetic ratio |
| $B_0$ | Magnetic field |
| $S$ | Signal intensity |
| $tanh$ | Hyperbolic tangent function |
| $y$ | True label/Ground truth |
| $\frac{d}{dx}$ | Derivation by $x$ |
| $p$ | Predicted probability |
| $o$ | Output |
| $DSC$ | Dice similarity coefficient |
| $G_{ss}$ | Slice selection gradient |
| $L_n$ | binary cross-entropy loss |
| $z$ | Loss |
| $f$ | Function |
| $K$ | Two-dimensional kernel |
| $I$ | Two-dimensional input image |
| $\cap$ | Intersection |
| $\pi$ | Pi |
| $\propto$ | Proportional |
| $\sum$ | Summation |
| $\infty$ | Infinity |
| $e$ | Exponential function |

# 1. INTRODUCTION

Lower back pain (LBP) is a common health problem around the world [1, 2]. The National Institute of Neurological Disorders and Stroke (NINDS), considers LBP as one of the leading problem of job-related disability worldwide [3].

Magnetic resonance imaging (MRI) is one of the imaging methods used by clinicians to identify pathoanatomic changes in LBP patients, as scans create detailed images of the spine. MRI is commonly requested to image the spine for patients with LBP and in the case of most spinal diseases as it can visualize changes in the structure of the spine revealing potential injuries [4, 5]. $T_1$-weighted and $T_2$-weighted imaging sequences are the two most commonly used contrast weightings in MRI. $T_1$-weighted spine imaging is used to identify fatty tissue and to obtain morphological information where fluid is visualized as dark regions [6]. On the other hand, $T_2$-weighted spine imaging is used to detect for example inflammation and edema, where fluid is visualized as bright areas [6].

After the imaging phase, a radiologist will examine the produced images and give a diagnosis. Due to the increasing amount of radiological examinations and overload in terms of time and resources, artificial intelligence (AI) enhanced analysis could be beneficial. The rapid development of deep learning (DL) methods especially convolutional Neural Networks (CNNs) [7], gave it superiority to overcome different tasks and problems related to the medical imaging field such as detecting and segmenting tumors and detecting patterns from data ([8]). One example of such an AI model is SpineNet [9]. SpineNet was developed to assess spine degeneration in MR images. The model is used to assess Pfirrmann grading, Modic changes, disc narrowing, spondylolisthesis, and central canal stenosis, and produce and locate radiological gradings automatically from spinal lumbar MRIs with near human performance level results [9]. In a population-based study, radiomics texture analysis was employed on the Northern Finland Birth Cohort 1966 data showing that LBP can be attributed to texture statistics in $T_2$-weighted MRI ([10]). If the process of detecting LBP could be automated, it might substantially help with the overload issues mentioned earlier by shortening the time required for image assessment. In this work, the applicability of CNNs to LBP classification from lumbar MRI data was investigated.

Finding the cause of LBP is a very challenging task and many studies have been conducted to find a connection between MRI findings and LBP [11, 12, 10]. AI-based LBP assessment could have the potential to enhance and aid medical doctors in the diagnostic process.

The aim of the first part of this thesis was to develop a method to automatically crop the relevant area containing lumbar vertebrae and intervertebral discs from lumbar spine MRI. DL-based segmentation method was developed using hand-drawn masks where the lumbar spine region was automatically segmented in for the rest of the (unmasked) images. The aim of the second part of this study was to predict the presence of LBP by analyzing the segmented $T_2$-weighted lumbar spine MRI images using a CNN model. Furthermore, we wanted to compare predictive ability by using different inputs from the MRI studies: i) The mid-sagittal slice from each MRI study, ii) The mid-sagittal slice images $\pm 4$ mm (3 slices), and iii) mid-sagittal slice images $\pm 4$ mm and $\pm 8$ mm (5 slices).

# 2. BACKGROUND

## 2.1. Lumbar Spine

The spine is a flexible column of bone, that extends from the skull to the coccyx (tailbone). It consists of 24 individual vertebrae bones, and two sections of fused vertebrae (sacrum and coccyx) located at the bottom end of the spine. The spine is divided into 5 sections: **1) cervical:** consists of 7 vertebrae of the neck (C1-C7), **2) thoracic:** includes 12 vertebrae (T1-T12), **3) lumbar:** consists of 5 vertebrae (L1-L5), **4) sacral:** consist of 5 fused vertebrae, and **5) coccyx:** consist 4 fused vertebrae (Figure 1) [13]. The spine provides structural support for the body and the nervous system.

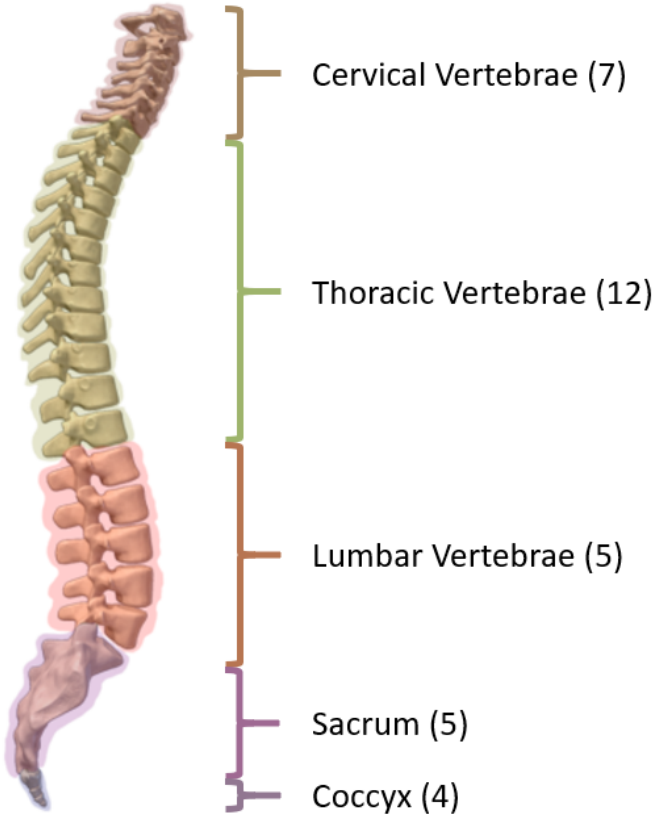
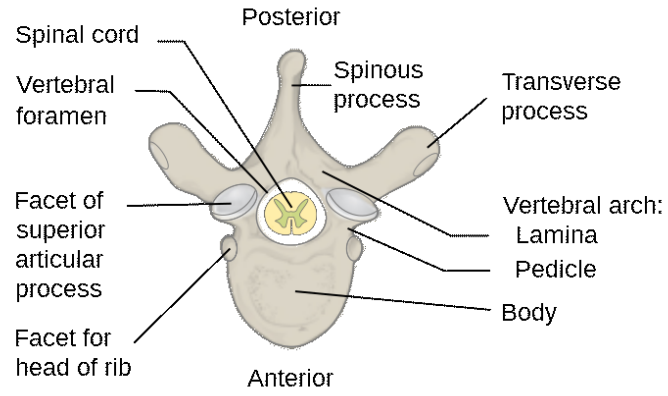



Figure 1. Spine anatomy showing segments of the vertebral column with the names and number of vertebrae. (Figure was reprinted [modified] from [13] under the Creative Commons Attribution-Share Alike 4.0 International license.)

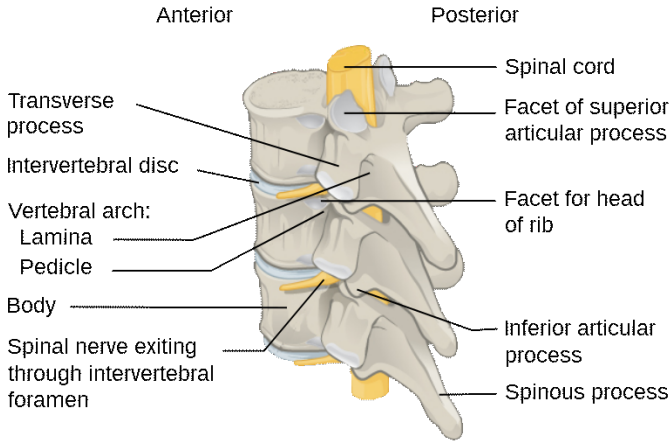
Besides bones, the spine relies on other supporting structures such as intervertebral discs (IVDs), facet joints, and spinal ligaments [14]. The gaps between the vertebrae are sustained by IVDs that act as inter-body spacers and shock absorbers in the spine when the body moves. At the back of each vertebral body (C3-L5) are the facet joints that help stabilize the spine and at the same time allow twist, extension, and flexion movement [14]. The vertebrae are held in place by ligaments and tendons that connect

the vertebrae and the muscles to the spine. The ligaments allow the bones, discs, and joints to bend, flex and twist within a limited range [14].

The spine provides protection to the spinal cord that is located within the vertebral canal (Figure 2 (a)) [15], and nerve roots (Figure 2 (b)) [16]. The spinal cord and nerve roots are part of the central nervous system that control body movements and transmit signals from the body to the brain. In this work, the focus is set on the lumbar spine which consists of the vertebrae L1 to L5 and IVDs L1-L2 to L5-S1.




(a)




(b)

Figure 2. Examples of (a) the superior view and (b) the lateral view of IVDs. (Both figures were reprinted ((a) from [15] and (b) from [16]) under the Creative Commons Attribution-Share Alike 3.0 Unported license.)

### 2.2. Low Back Pain

Low back pain (LBP) is a symptom caused by a variety of syndromes taking place in the lumbar spine (Figure 1) and surrounding musculature and other tissues [17, 18]. The pain may result from injury or degenerative processes of the lumbar spine innervated tissues, IVDs, ligaments, or muscles. Signs of disc degeneration may lead

to LBP [19]. The pain can be also defined as muscle tension or stiffness localized below the costal margin and above the inferior gluteal folds [17]. Usually, LBP can be categorized into specific or non-specific pain as it depends on the symptoms and the spinal pathology [18]. Specific LBP in the case where there are symptoms that are caused by pathophysiological changes like inflammation [20], infection, rheumatoid arthritis, osteoporosis, fracture, or tumor [21]. In some cases LBP can become chronic, and can be a disabling condition in extreme cases [17]. When there are no pathological changes in clinical examinations, LBP is considered non-specific [21].

Damage to the IVDs can lead to herniated discs and that may lead to spinal pain (see for example [22]). The disc consists of hard and soft tissue where the outer layer (annulus fibrosus) is tough, and the center (nucleus pulposus) is a gel-like substance consisting mainly of water and collagen fibers. When the outer layer of the disc ruptures, the substance in the center may leak and towards the spinal canal and press against the nerves there leading to pain strikes. The pain may be also accompanied by other symptoms such as numbness, and weakness in an arm or a leg [22].

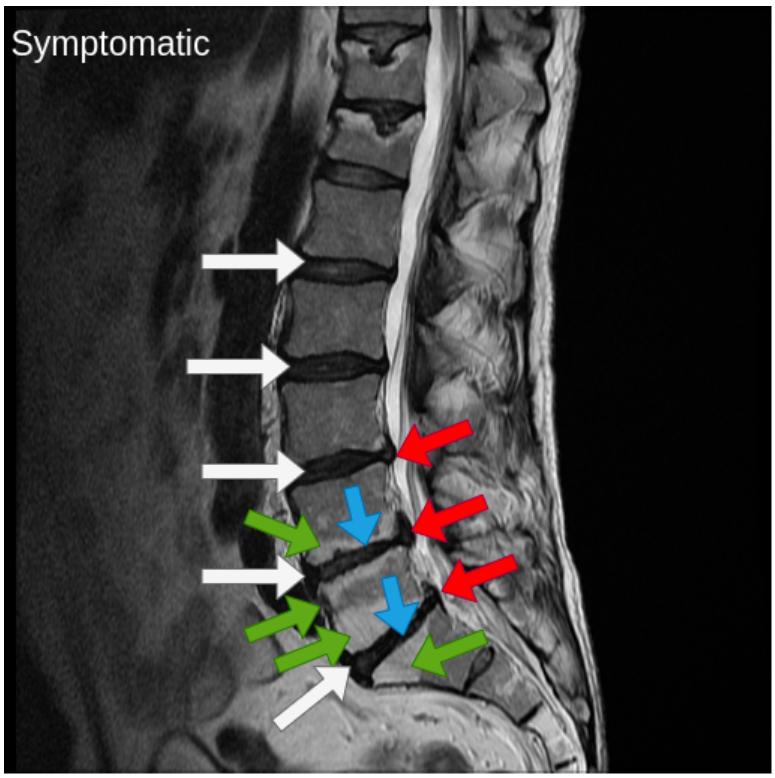
Many other conditions, diseases, and injuries may lead to LBP such as structural problems, spondylolisthesis, disk problems, strains and sprains, fractures, and other diseases. Structural problems such as spinal stenosis accrue when the spine is narrow for the spinal cord leading to pain. Spinal osteoarthritis is a common type of arthritis that can lead to LBP [23]. Accidents can lead to fractures in the spine thus leading to LBP, and in some pathological cases such as spondylolysis or osteoporosis, the risk of fractures is higher [24]. Lower back strains and sprains are common causes of LBP, as it is relatively easy to injure ligaments, muscles, or tendons by lifting heavy items or doing physical activity. Diseases such as infections, spine tumors, and cancer can cause LBP. Literature has shown that disc height narrowing, joint degeneration and endplate lesions such as Schmorl's nodes, erosion, fractures, calcifications, and annular tears are all related to LBP [25].

### *2.2.1. Economical Load*

LBP is one of the burdensome diseases worldwide, and it is associated with high economic costs as it is related to healthcare costs and lost productivity worldwide [26]. In the US, LBP is considered a common cause of wasted work time, and the second most common cause of disability [27]. In the US alone, it is estimated that 149 million days of work are lost due to LBP per year [27]. The healthcare spending in the US on LBP is estimated to exceed $ 100 billion in 2005 [28], and in 2013 it was the third-highest of healthcare spending at 87.6 billion [29]. In Sweden, the economic load of LBP for all episodes that started in 2011 was estimated to be at 740 million euros [30].

### *2.2.2. Diagnosis*

Pain is subjective and different from one person to another, thus it is challenging to specify the cause of pain. Medical professionals usually use the input of patients about the symptoms and do a physical exam to diagnose the cause of pain. In some cases, imaging is needed in order to check for broken bones in the spine and other damage




(a)


(b)

Figure 3. Examples of the middle slices from the data showing both (a) a symptomatic and (b) an asymptomatic subject with colored signs of spinal degeneration. Arrows correspond to herniation (red), disc degeneration (white), endplate changes (green), and collapsed disc (blue).

or degeneration in tissues around it. Imaging methods such as spine X-ray, magnetic resonance imaging (MRI), and computed tomography (CT) scans are usually used. MRI is mainly used to visualize soft tissue, whereas CT and X-ray are (mainly) used to visualize the vertebrae (e.g. in trauma cases). CT has limited soft tissue contrast and limited ability to visualize the properties of the disks ([18]). MRI is the preferred method to investigate soft tissue structures and detect nerve root entrapment, lumbar disc herniation and spinal canal stenosis [11].

In other cases, blood tests or urine tests are used to detect genetic markers for some conditions such as ankylosing spondylitis that are related to LBP [21]. Diseases such as polycystic kidney disease may lead to pain in the low back making it look like LBP [31].

The degenerative changes in vertebral endplates and bone marrow observed in $T_1$- and $T_2$-weighted MR images that are defined as Modic changes that can be categorized into three main types (Table 1) [32]. Literature has shown that Modic changes are associated with LBP [33]. IVDs degeneration can be also assessed with the five grade using the Pfirrmann scoring method from $T_2$-weighted MR images (Table 1) [34].

Table 1. Pfirrmann grading system and Modic changes classification with a description of how changes are graded from MR images [34, 32].

| Classification | Description |
|---|---|
| Pfirrmann 0 | The disc is normal and no degeneration signs are observed. |
| Pfirrmann 1 | Disc structure is homogeneous with a bright hyperintense white signal intensity and disc height are normal. |
| Pfirrmann 2 | Disc structure is inhomogeneous, with a hyperintense white signal. |
| Pfirrmann 3 | Disc structure is inhomogeneous, with intermediate gray signal intensity. |
| Pfirrmann 4 | Disc structure is inhomogeneous, and a hypointense dark gray signal intensity. |
| Pfirrmann 5 | Disc structure is inhomogeneous, with a hypointense black signal intensity. |
| Modic Type-0 | Disc and vertebral body have no degenerative changes. |
| Modic Type-1 | Bone marrow edema is present within the vertebral body and hyper-vascularization. |
| Modic Type-2 | Fatty replacements of the red bone marrow appear within the vertebral body. |
| Modic Type-3 | Bone sclerosis is visible. |

### *2.2.3. Management and Treatment*

Treatments for LBP are different and depend on the cause of the pain, including medications, physical therapy, injections, and surgery. Most LBP is acute or non-specific which gets better in time, with rest, ice, anti-inflammatory drugs, muscle relaxation medications, and pain relievers if needed [21]. Resting for a few days after injury can be enough in most cases to get back to normal life activities.

Physical therapy leads to improvement in flexibility, posture, and alignment which works on supporting the spine by strengthening the muscles around it and can take the pressure from the spine. Increasing the blood flow by doing physical activities can help

in the healing process [35]. The increase in blood flow increases the nutrients input to the soft tissues in the low back increasing mobility and improving healing. Aerobic exercise is one of the methods used to ease chronic LBP as the body will produce more endorphins after doing 30-40 minutes of exercise which is like taking pain-reducing drugs such as codeine or morphine [36].

Treatment by injections is done usually by injecting epidural steroids into the area that causes the pain to reduce inflammation thus relieving the pain [37]. Epidural steroid injection is beneficial for short-term acute episodes of LBP, but injections are not always effective. In some rare conditions, surgical solutions are needed to repair the spine and help patients to live pain-free life. Lumbar spinal fusion is one option used for deformities of the spine, replacing the herniated disks, and providing more stability to a weak spine, as two or more vertebral bones fused permanently together by replacing the soft tissues between them with bone or metal and letting them heal to one solid vertebral bone [38]. Literature shows that lumbar spinal fusion is not the most effective treatment for chronic LBP [39]. Lumbar artificial disc replacement surgery is another option that is a potential alternative to fusion surgery with a faster recovery time and the artificial disc provides more flexibility [40]. Non-specific LBP surgical treatments, in general, are not recommended [41].

### 2.3. Magnetic Resonance Imaging with $T_2$-Weighted Contrast

MRI is an imaging methods based on the nuclear magnetic resonance (NMR) phenomenon. It produces a detailed image of the lumbar spine region and the tissues surrounding it [42]. MRI can be used to visualize soft tissues with different contrast weightings that determine the contrast and brightness of the image. The use of MRI can visualize other spinal structures such as vertebrae, IVDs, and the tissues around it [18].

NRM is a physical phenomenon in which atomic nuclei placed in a strong magnetic field absorb Radiofrequency (RF) electromagnetic radiation at a resonance frequency $\nu$ [43]. The resonant frequency is proportional to the strength of the magnetic field $B_0$ acting on the nuclei as follows:

$$\nu = \frac{\gamma B_0}{2\pi}, \tag{1}$$

where $\gamma$ is the gyromagnetic ratio of the nucleus. Paul Lauterbur produced the first NMR image in 1971, by using the magnetic field gradients to encode spatial information into an NMR signal [44]. In MRI, the imaging views can be obtained from any orientation, and in order to do that the images must be obtained as separate imaging sequences with a set of parameters for each sequence making imaging periods longer when more sequences are done. The sagittal plane is considered the best view to define spine pathology [45].

The MRI device consists of a large tube containing a large magnet, gradient coils, radio frequency coils, receiving coil, a bore, and a table where the patient is positioned during imaging (Figure 4). The patient needs to stay still for image acquisition in order to have good-quality images without motion artifacts.

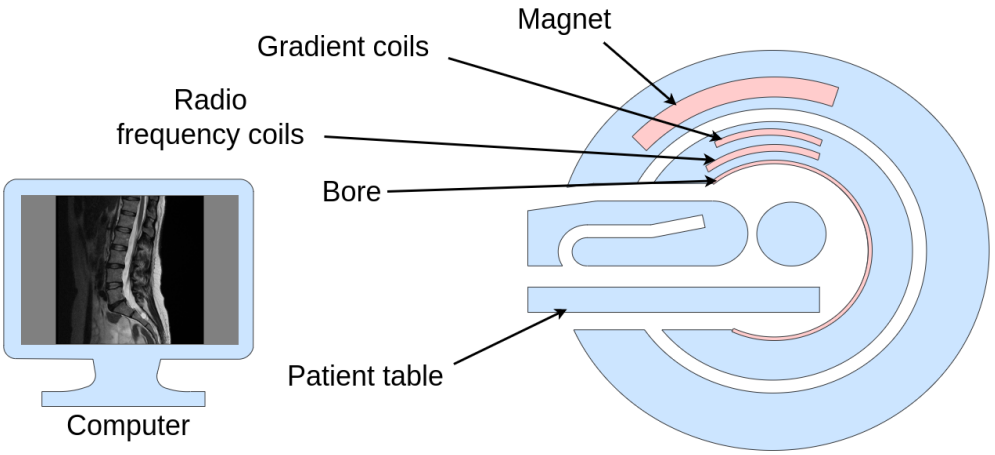



Figure 4. Example of an MRI scanner and its main components.

The magnet inside the MRI machine tube generates a strong magnetic field ($B_0$) making the protons inside the body line up with the magnetic field [43]. Specifically, the spinning nuclei will start to precess around the $B_0$ axis as they realign and changes the spin direction of atoms in the body temporarily. Spin is one of the four physical properties (spin of elementary particles causes a magnetic moment) of the atom on which NMR is based. In general, most atoms have spin properties where thermal energy makes the atoms rotate about their axis creating angular momentum. The spin is pointed in all directions in the absence of an external magnetic field. The signal in MRI comes from hydrogen in water and fat that have spin angular momentum, giving the nuclei a magnetic moment. Water masks up $70\%$-$90\%$ of most tissues consist and each water molecules have two hydrogen atom.

In order to obtain MRI signal, first, the transmitting coil transmits RF pulses at the proton/hydrogen NMR frequency. RF pulses are applied to tip the spin direction away from the external magnetic field ($B_0$) direction. The flip angle is the change in spin direction from the external magnetic field with the help of RF pulses. When stopping the RF pulse, the protons return back to the previous spin orientation (relaxation). This setup is repeated and the emitted free induction decay (FID) signal is measured where frequency and decoding are used to obtain the spatial information.

The used pulse sequence defines echo formation, slice selection, position and frequency, and the readout gradient which are divided into either spin echo or gradient echo. In the spin echo (SE) method, first, a 90 degree excitation pulse (see figure 5) is emitted, which causes the NMR phenomenon where the nucleus absorbs the energy of the pulse and the nuclei turn (flip) 90 degrees in the direction of transverse magnetization. The formation of echo depends on the sequence used, which is divided into either spin echo or gradient echo techniques (Figure 5). After excitation, the spins naturally dephase until at time TE/2 (where TE stands for time of echo) a new refocusing 180-degree pulse is sent which reverses the phases of the spins, and thus at the TE the echo is formed and the signal can be measured with the receiving coil, which decays exponentially with the spin

$$S_{SE} \propto S_0 e^{-TE/T_2}, \tag{2}$$

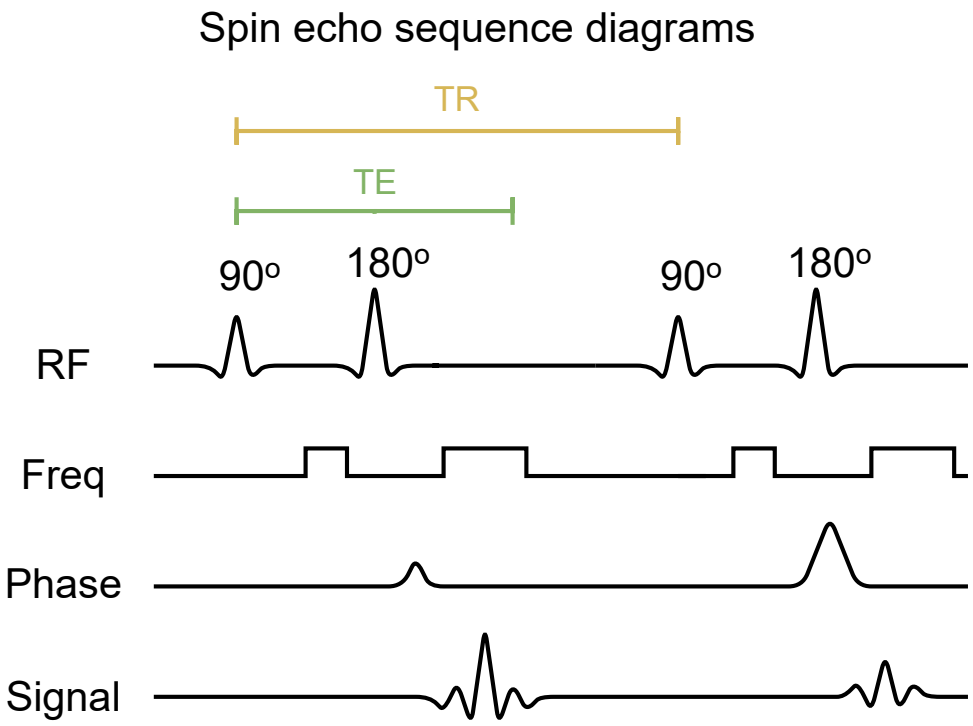



Figure 5. Example of transmitted spin echo sequences. Radiofrequency (RF) pulse, frequency encode gradients (Freq), phase encoding gradients (Phase), the MR signal (Signal), time of echo (TE), and time of repetition (TR).

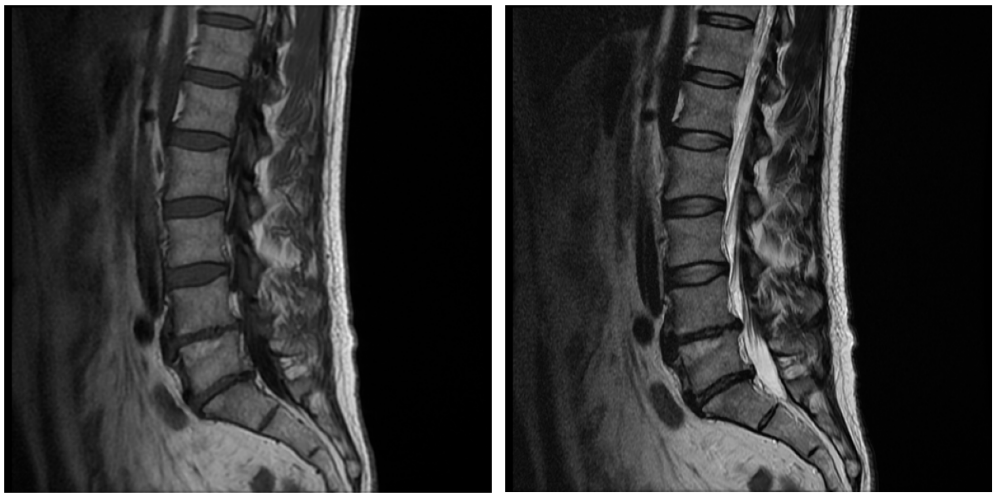
where $S$ is the signal intensity.

$T_1$-relaxation (also known as longitudinal relaxation) and $T_2$-relaxation (also known as transverse relaxation) weighted scans are two common MRI contrasts (Figure 6). $T_1$-weighted images are useful for identifying fatty tissue and obtaining morphological information, where fluid in this sequence is dark [6]. $T_2$-weighted image is used to detect edema and inflammation, revealing white matter lesions where fluid in this sequence is bright ([6]). The reconstructed image contrast whether it is $T_1$- or $T_2$-weighted depends on the parameters (such as TR, time to echo TE, and RF pulse) and sequence (such as SE and GE) during imaging. Both $T_1$ and $T_2$ relaxation occurs using the same sequences. The parameters of the sequences can be used to influence which (or both) effect dominates the signal, and where the contrast weighting comes from. $T_1$-weighted needs short Repetition Time (TR) and TE whereas $T_2$-weighted needs long TR and TE depending on which sequence is used.

The change in magnetization direction (toward and away from the coil), and induction creates FID which is the current in the coil (MRI signal). Gradient coils are used to change the magnetic field and thus the resonance frequency. Gradient coils are needed to encode the spatial information of the image. When the 90-degree RF pulse is sent, at the same time the MRI device turns on the slice selection gradient ($G_{ss}$) that consists of the sum of $G_x$, $G_y$, and $G_z$ gradients. The MRI device can select any direction point by combining the gradients $G_x$, $G_y$, and $G_z$ where the wanted slice is coded with gradients. Slice selection gradients and RF pulses are used to select image slices and the gradient can be used to locally modify the strength of the magnetic field in the wanted area so that the RF pulse excites a certain slice of the person. The

frequency encoding uses additional magnetic field gradients in order to achieve spatial positioning. After this, a Phase encoding gradient is applied, which is perpendicular to the frequency encoding gradient. When the echo is measured in TE, the encoding gradient of the frequency direction is turned on and the signal is measured with the receiving coil. This way, one line of measured signal is collected in the K-space matrix. After the K-space has been filled, the inverse 2D Fourier transformation is applied to convert the signals to gray levels in the cross-sectional MRI image in cross-sectional MRI images.

Vertebral endplates and IVDs can be clearly visualized in $T_2$-weighted MRIs [46], and fat infiltration which is strongly associated with LBP, is also visible in $T_2$-weighted MRI [12]. From $T_2$-weighted images, IVDs degeneration which is associated with LBP can be assessed with the five-grade Pfirrmann scoring method, as signal loss and nonuniformity of the disc correlates with the progressive degenerative changes [34, 19, 47]. Modic changes can be observed by comparing signal intensities between $T_1$ and $T_2$ -weighted MRIs ([48]). Disc herniation, which can help in the diagnosis of sciatica, can also be seen in $T_2$-weighted MRI ([49]).



(a) (b)

Figure 6. Example of (a) $T_1$-weighted where fatty tissues are bright, the spinal cord is gray, and disks are gray (b) $T_2$-weighted images where fatty tissue and water are bright, the spinal cord is light gray, and disks are bright.

## 2.4. Deep Learning

Deep learning (DL) is a sub-field of machine learning (ML) and artificial intelligence (AI), that imitates the way humans gain knowledge by learning from raw data (see for example [50]). DL was first invented in Ukraine by Alexey Ivakhnenko in 1965 [51]. Convolutional neural networks (CNN) [7] are one of the most popular types of neural networks, with excellent performance in machine learning problems, especially in computer vision applications.

### *2.4.1. Deep Learning in Medical Image Analysis*

DL has shown superiority in solving many medical image-related problems such as classifying and segmenting pathology images with tumor regions and/or metastasis detection [52]. DL has a lot of challenges, especially ones related to medical image analysis such as learning from faulty datasets, shortage of properly annotated labels, imbalanced datasets, and where DL may misinterpret input data by focusing on the wrong features [53, 54].

The quality of the medical data is considered one of the main challenges in DL. In general, the amount of data effect directly on the performance of the model as the DL model needs large datasets to learn to do more complicated tasks. The selection of the suitable DL model architecture that takes into account the data available is also one of the DL challenges [55]. Imbalanced data add yet another challenge to DL data analysis [56].

The use of ML and its sub-field DL has seen exponential growth in the analysis of medical data [57, 58]. ML has been applied successfully to diagnosis (pediatric LBP, normal and abnormal cervical spine vertebra, deformity of scoliosis spinal, and posterior lumbar spine fusion risk factors) and prognosis of the outcome of the complications after spine fusion surgery, LBP, and other spinal diseases [57].

### *2.4.2. Artificial Neural Networks*

Artificial neural networks (ANNs) are bio-inspired systems based on studies of the human brain and the nervous system [59]. The principles of which the ANNs work is based on forward propagation in order to use the input to generate the output (Figure 7) and backward propagation where the error is backpropagated from the output to the input. The backward propagation is started by a loss function where the ANN model learns when the input data will go back and forth using forward and backward propagation. Neurons are the building block of ANNs architecture, consisting of a summation function, weights, and an activation function used to process data (input) to get the wanted output. The summation function is followed by an activation function that converts the weighted sum of inputs to an output (forward propagation). There are many types of activation functions such as the sigmoid function which is used usually at the output layer of binary classification and is defined as

$$f(x) = \frac{1}{1 + e^{-x}}, \tag{3}$$

the hyperbolic tangent function which can be used in hidden layers and is defined as

$$tanh(x) = \frac{2}{1 + e^{-2x}} - 1, \tag{4}$$

and the Rectified Linear Unit (ReLU) activation function which is used in almost all convolutional neural networks, and is defined as

$$f(x) = max(0, x), \tag{5}$$

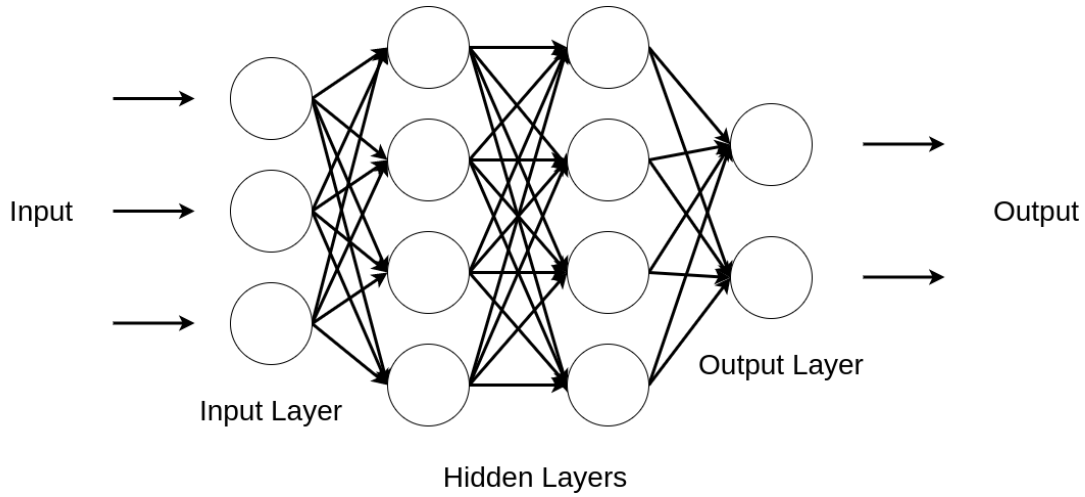



Figure 7. Example of a simple fully connected ANNs with an input, an output, and 2 hidden layers.

where $x$ is the neuron input. In backward propagation (Backpropagation), the gradient of the loss function for each weight is calculated based on the error in the previous epoch using the chain rule method. The chain rule is defined as

$$\frac{dz}{dx} = \frac{dz}{do} \cdot \frac{do}{dx}, \tag{6}$$

where $d$ refers to the derivative of the input $x$, the loss $z$, and the output $o$. The chain rule computes neuron weights in each layer by updating them based on the error in ANNs making the models more reliable. Backpropagation is performed after each forward propagation. Modern backpropagation was developed in Finland in the 1960s by Finnish master's student Seppo Linnainmaa [60].

ANNs are designed to obtain the statistical generalization of a function, as the neurons that contain activation functions are approximation engines guided by mathematical and engineering domains [50]. ANNs are powerful models able to automatically extract patterns from large datasets [61]. Usually, ANNs have hundreds or thousands of interconnected artificial neurons allowing them to be powerful models (Figure 7).

### *2.4.3. Convolutional Neural Networks and Residual Neural Networks*

Convolutional neural network (CNN) is considered one of the most used models in the DL field (see for example [62]). Convolution [50] is described as

$$S(i,j) = (K * I)(i,j) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} I(i-m, j-n)K(m,n), \tag{7}$$

where $I$ refers to the two-dimensional input image and $K$ to the two-dimensional kernel. CNN has the potential to learn complex patterns present in a set of images. CNN uses randomly initialized filters in order to extract features from the input images

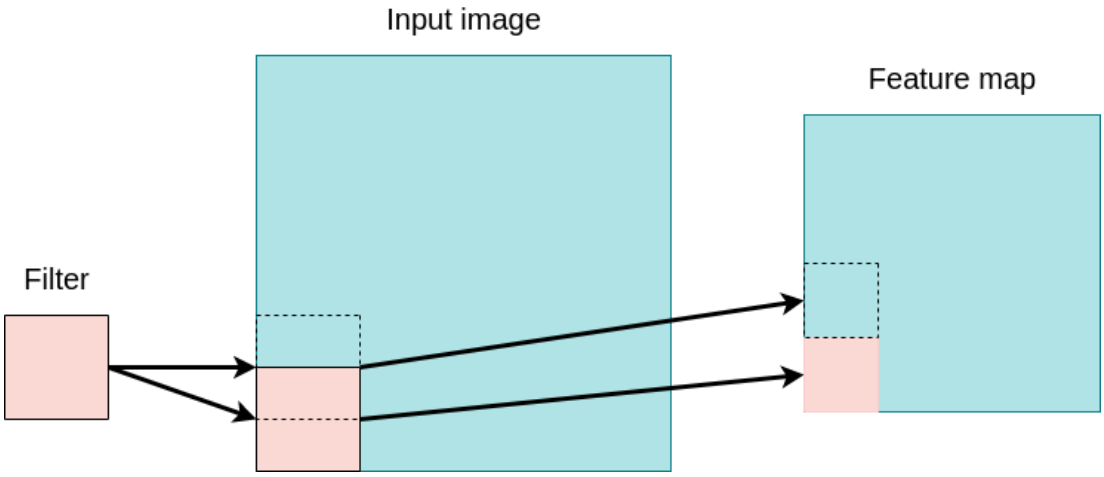



Figure 8. Example of filter (kernel) applied to an input image to construct a feature map.

thus CNN models learn the best filters by updating the filter weights. CNN overall consists of two main functions, the input image and the filter (kernel). Each point (pixel) of the image is multiplied with the filter elements highlighting the patterns. The CNN algorithm may have many layers where each layer can have a different size filter and a number of strides desired. The number of hops the filter makes across each image is known as strides, and each layer may have any filter size and number of strides needed. The filter shifts as it moves across the image's pixels creating an image that contains the highlighted features. The created features depend mostly on the filter's structure. This process is repeated at each layer forming gradually more abstract features. The output image generated from the CNN is also known as a feature map (Figure 8). CNN architecture may also include other types of networks, such as attention models or fully connected.

After 2012 when CNN-based architecture AlexNet [63] won the ImageNet competition, CNN models were made deeper and deeper, but the performance worsened due to the vanishing gradient phenomenon. In 2015 residual neural networks (ResNets) was introduced to solve this problem [64]. ResNets allows DL models to be much deeper and train more effectively to solve complected problems. The network uses skip connections to connect activations of different layers together and skip some layers in between. This connection allows the network to skip any layer that may damage the performance. As a result of this, the recent architecture allows the training of very deep neural networks without facing vanishing gradient problems.

### *2.4.4. Supervised and Unsupervised Learning*

Supervised learning is an AI approach of creating an algorithm that is trained on input data using certain labels in order to output data as a reference for the training data [65]. For example, images of dogs and cats can be used to teach a CNN model how to classify them automatically (Figure 10) [66]. Models are trained to detect the patterns and connections between the input and the output data, making them able accurately label new unseen data. Adaptive moment estimation (Adam) optimization algorithm

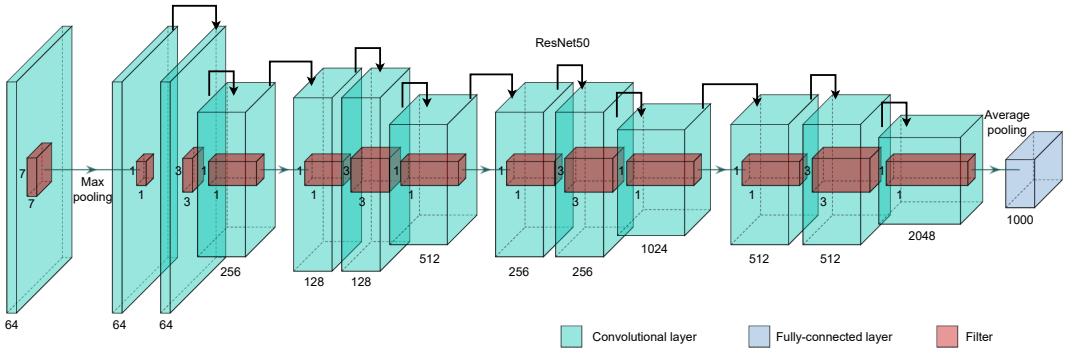



Figure 9. Pre-trained Resnet-50 architecture representation. The arrows on top of the convolutional neural networks stand for the skip connections.

can be used to optimize the model [67]. On the other hand, unsupervised learning is the approach where the algorithm is presented with unlabeled data and it is designed to detect patterns or similarities from the input data on its own. During each iteration, all the weights are updated proportionally to the partial derivative of the error function with respect to the existing weights. When the number of layers is small, this can be a very effective technique.

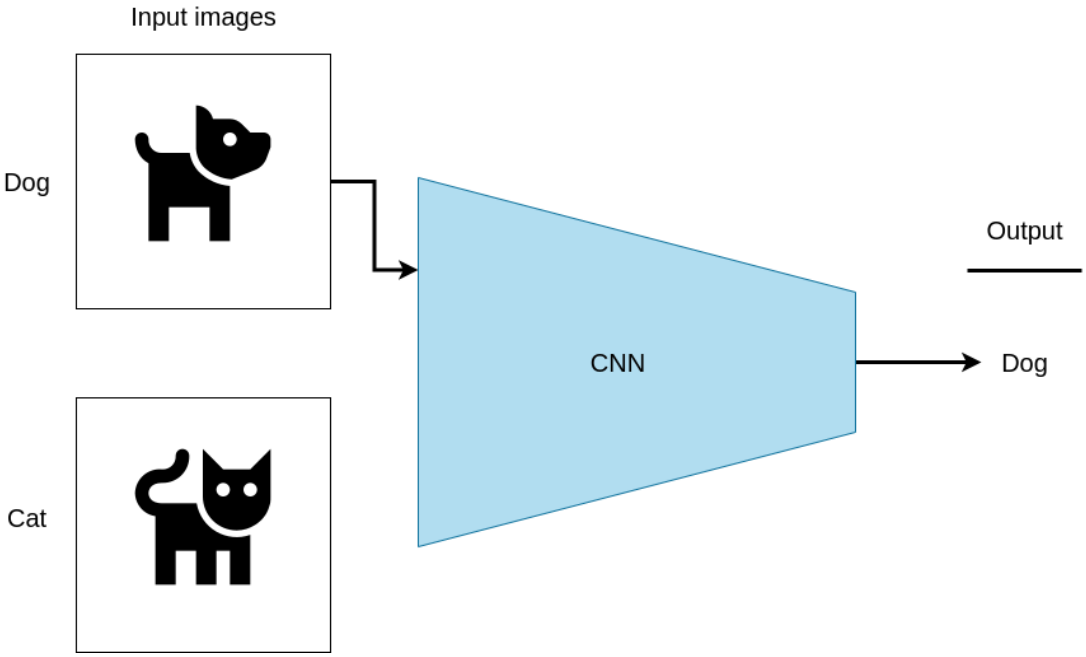



Figure 10. Dogs vs cats classification example using CNN models.

### *2.4.5. Optimization*

Optimization in general refers to minimizing a specific loss function that optimizes the performance of a given method [50 , Chapter 8]. DL algorithms usually rely on some optimization method such as stochastic gradient descent or Adam [68, 67]. The

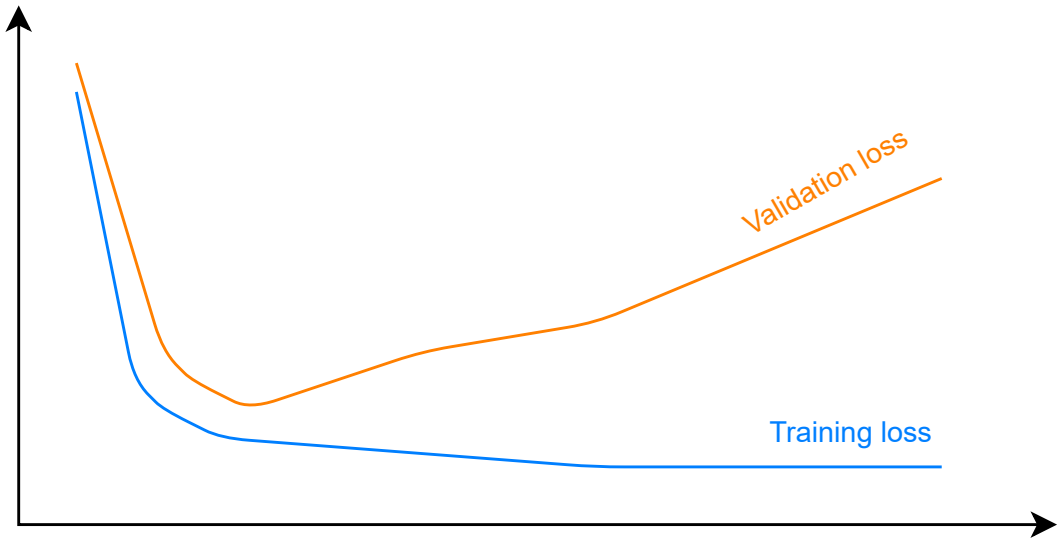



Figure 11. Example showing how overfitting looks like from training and validation losses. Overfitting happens when the loss starts to increase after it was decreasing along with the training loss.

loss function is used to provide feedback on the performance of the DL model by measuring the errors made by a neural network [50]. For example, loss functions like cross-entropy loss [69] and binary cross-entropy loss [7] are used to optimize the performance of the neural network. The loss function is one of the essential components in order to train DL models more efficiently. This impact makes loss function selection quite critical, as different loss functions may generate different results in the same network's performance.

In cases where the dataset is unbalanced (one class is more represented than the other) weighted loss functions such as weighted binary cross-entropy [70] can be used to remove bias.

### *2.4.6. Overfitting and Cross-Validation*

The main purpose of the training process is to maximize the accuracy of the prediction of new unseen data by the model. The dataset is usually split to train, validation, and test (hold out) datasets. The training dataset is used to train the model, where the results of the validation dataset will be used to set the hyperparameters. Finally, the test (hold out) dataset will be used to evaluate how the model is performing to new unseen data. Overtraining or overfitting is a drawback that may take place when the training model learns to perform very well on the training data and fail to do so on new data [50]. To check if the model is more generalized and not overfitting can be done by comparing the training and validation datasets losses (Figure 11). DL model is considered overfitting when the training loss value will continue to decrease while validation loss will increase at some point (Figure 11).

Cross-validation is a resampling method where different parts of the dataset train and validate DL models on different iterations. Cross-validation is mainly used to test

the performance of machine learning models in practice [65]. It is one of the used statistical techniques to overcome overfitting in a predictive model, especially in some cases where the amount of data is limited. The training data is split into two parts in each iteration, where one part is used to train the model and the other is to validate the learning process. This procedure has only one parameter called $k$, and it refers to the number of folds that a given dataset is to be split into. This way cross-validation is sometimes called k-fold cross-validation (Figure 12). After specifying the number of folds (k), the data will be split into $k$ folds, the model is run $k-1$ of them, and validated on the remaining part. The average of each validation performance is calculated after such runs (Figure 12).

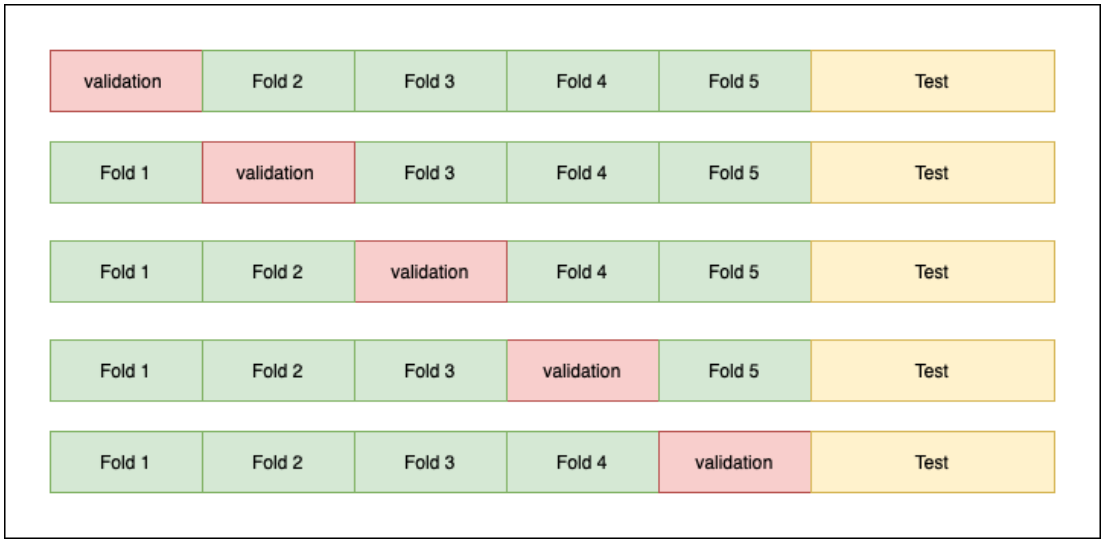



Figure 12. Different folds used in 5-folds cross-validation setting.

### *2.4.7. Regularization Methods*

The purpose of regularization is to avoid overfitting [50]. In other words, regularization of DL models means how to make the models perform better on the new unseen data. Regularization is considered among the main methods related to DL [50]. To prevent overfitting, techniques such as weight decay, dropout, data augmentation, and cross-validation are used. Weight decay is weight adding to the loss function to decrease complexity and regularize the model. In order to implement weight decay, squares of all the parameters are added to the loss function. The values of the parameters could get too large after squaring, in relation to the main loss function. To prevent that, the sum of squares is multiplied with the weight decay value to decrease it making it smaller [71]. The weight decay value is a hyperparameter where too small values model can well fit the training data but test accuracy would be very low, and if the value is too high then the model will be restricted as will be forced to use extremely small weights making it not able to fit both training and test data. Weight decay is considered one of the standard methods for regularizing DL models [71].

Dropout is the process of removing some of the nodes in the hidden layers of neural network. The predictions are impacted greatly if the weights in some parts of the

network are too big than others. If the weights are too big in some parts, the model will overfit. In order to overcome this issue, the learned representations are distributed more evenly across multiple weights using dropouts so that the model does not overfit [72].

Another technique used to regularize DL models during training is data augmentation as it can prevent overfitting [73]. Data augmentation is based on the artificial generation of existing image samples from the training dataset randomly. The augmented images are altered through different ways of processing, such as flips, noise, random rotation, shifts, etc. Image augmentation is used to enlarge the existing dataset and also solve overfitting problems [73]. When trained on a large dataset, DL models tend to give better results [50, 71]. Using augmentation to make various versions of the images in the dataset could improve the model and its ability to generalize the prediction to new unseen images [73].

#### *2.4.8. Transfer Learning*

Transfer learning means the use of model knowledge that is previously trained (pre-trained) usually on a large dataset to solve a new but related task. Pre-trained models can be used as is, or transfer learning is used to modify the model to a new task. There have been also efforts to assemble large datasets consisting of labeled images such as ImageNet [74] in order to pre-train DL models such as ResNet [64], VGG [75], and AlexNet [76]. ML models that are trained to extract informative features for a specific task can be used as a starting point for a second model. In cases of a small dataset, DL pre-trained models are a popular approach as they increase the performance [77].

### 2.5. Image Segmentation

Image segmentation is the process of clustering and localizing the pixels of objects and boundaries in images. This kind of clustering and localizing transform digital images into various classes, making it easier for computers to analyze them [78]. Image segmentation methods output a set of segments or contours covering parts of the image that belong to the same object class. In computer vision and digital image processing image segmentation is a popular problem [78]. Further analyses such as object detection are possible after segmenting and clustering part of the image [78]. Regarding medical imaging, clinically relevant information could be extracted from images using image segmentation [79].

U-Net is a CNN architecture used for semantic segmentation (Figure 13) [80]. The network was originally developed to segment medical images. The network architecture contains two paths: encoder (contraction) path and decoder (expansion) path. The encoder consists of convolutional and max pooling layers used to learn the features of the context in the input image. The convolutional layer applies a filter to the input image which outputs a map of activations also called a feature map. Max pooling is a function that pools the highest value from the feature maps, decreasing parameters in the network by reducing the size of the feature maps. The decoder is a symmetric expanding path where transposed convolutions are used to generate the

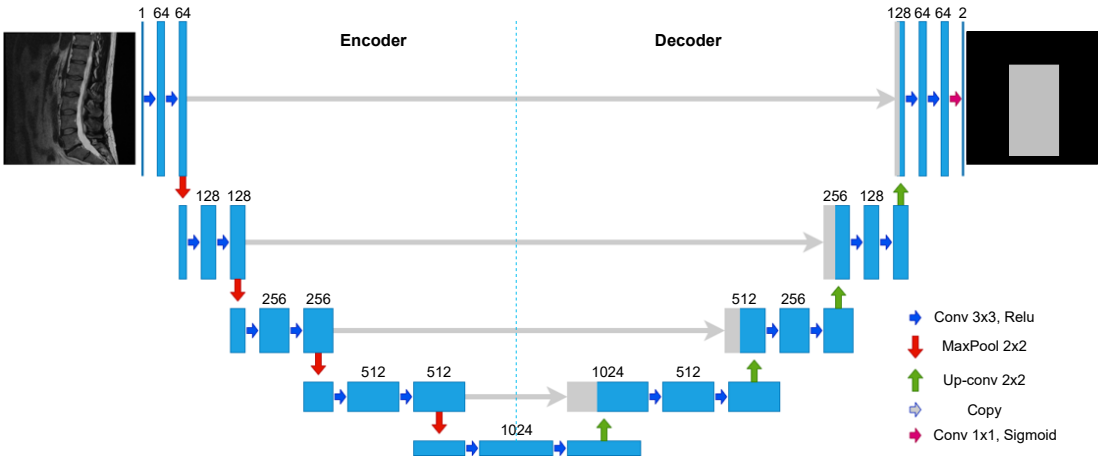



Figure 13. Example of the used U-Net network architecture for segmenting lumbar spine. The network used MR images and handmade masks to train in order to segment new MR images automatically. Conv stands for convolution where the number of the channels is denoted on top.

output image. Skip connections are used to concatenate the information (feature maps) further up the network. This will allow the information to propagate forward and help fix the vanishing gradients problem as it passes information through the network.

Several studies on medical imaging segmentation have been introduced in the past [81, 82, 83]. In the present work, DL is applied to crop the lumbar spine region from the middle slice of MR images (Figure 13).

## 2.6. Image Classification

Image classification is the task of assigning a label or class to an entire image. CNN has demonstrated outstanding results in computer vision tasks making it the most commonly used algorithm for image classification. DL advantages especially CNN architectures, made it well-suited for supervised and unsupervised diagnostic classification problems [84]. The ability of CNNs algorithms to be generalized in order to solve different tasks using similar designs is one of CNN's distinct advantages. For example, a CNN model such as ResNet [64] can be used for the classification of cats and dogs from images or to classify symptomatic and asymptomatic LBP cases from MRI images. DL networks have been developed and improved to achieve state-of-the-art accuracy on image classification tasks such as ResNet [64], VGG [75], AlexNet [76].

## 2.7. Evaluation Metrics

Evaluation metrics are used to quantitatively describe the behaviors of DL models as a part of validating their performance. Different metrics compare the results obtained from the model (predictions) with the actual labels. The metrics then generate scores based on that to show the effectiveness of the model. All evaluation functions used in

this study have a range of values between 0 and 1 (perfect prediction), except MCC where the range of values is between -1 (inverse prediction) to 1 (perfect prediction) (Table 2).

Table 2. The metrics that are used in the presented thesis. Where metrics are defined in terms of TP = True Positives (predicted as positive as was correct) FP = False Positives (predicted as positive but was incorrect), FN = False Negatives (failed to predict an object that was there), and TN = True Negatives (predicted as negative as was correct).

| Name | Equation | Range of values |
|---|---|---|
| Average Precision score | $AP = \sum_n (Recall_n - Recall_{n-1}) Precision_n$ | $[0, 1]$ |
| ROC AUC score | $ROC - AUC = \int_0^1 TPR(FPR)\, dFPR$ | $[0, 1]$ |
| Precision score | $Precision = \frac{TP}{TP+FP}$ | $[0, 1]$ |
| Negative predictive value | $NPV = \frac{TN}{TN+FN}$ | $[0, 1]$ |
| Recall score | $Recall = \frac{TP}{TP+FN}$ | $[0, 1]$ |
| Specificity score | $TNR = \frac{TN}{TN+FP}$ | $[0, 1]$ |
| Balanced accuracy score | $BACC = \frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{TN+FP})$ | $[0, 1]$ |
| Matthews correlation coefficient | $MCC = \frac{TN \cdot TP - FP \cdot FN}{\sqrt{(TN+FN)(FP+TP)(TN+FP)(FN+TP)}}$ | $[-1, 1]$ |
| False positive rate | $FPR = \frac{FP}{FP+TN}$ | $[0, 1]$ |
| True positive rate | $TPR = \frac{TP}{TP+FN}$ | $[0, 1]$ |

# 3. MATERIALS AND METHODS

## 3.1. Cohort Data

The Northern Finland Birth Cohort 1966 dataset (NFBC1966, http://www.oulu.fi/nfbc/) was used in this study which is a licensed biobank governed by the Biobank Act (688/2021). The collection of data began in 1965 in Northern Finland, as pregnant women living in Oulu or Lapland were expected to give birth between January 1st and December 31st, 1966, and were asked during their maternity clinic appointments to participate in the NFBC1966 birth cohort study. The health and lifestyle information on the mothers (*N*=12 068) and children (*N*=12 231) were collected, by postal questionnaires and clinical examinations ([85]).

Lifestyle and health questionnaires (*N*=10 321) were sent (2012–2014) to the cohort members with known addresses at the age of 46–48. More than half of the recipients (66%, *N*=6 825) responded to the questionnaires and clinical examinations were performed on the ones who were currently living in Finland which were 57% of the recipients (*N*=5 861). The questionnaire consisted of LBP-related questions, where the first question was "Have you had any aches or pains in your low back?", and a drawing to elucidate the correct anatomical region. There were two options to respond to the question, (1) no and (2) yes. In the case of a positive answer, the next question was "How often have you had aches or pains during the last 12 months?", and answers were (1) 1–7 days, (2) 8–30 days, (3) more than 30 days, and (4) daily. Members who reported pain in the past 12 months, were asked about the intensity of pain experienced using a Likert scale rating from 0 (no pain) to 10 (extremely severe or bothersome pain). The cohort members living within 100 km from the city of Oulu (*N*=1 988) were invited to undergo lumbar MRI examinations, and 1,540 underwent MRI of the spine. $T_2$-weighted images, where the imaging was done using 1.5 T MRI device in the sagittal direction. The images were acquired using a fast spin-echo sequence (Time of echo(TE) = 112.7 ms, Repetition time (TR) = 3 500 ms, Echo train length (ETL) = 27, slice thickness = 4 mm, matrix size = 512 $\times$ 512, and field of view (FOV) = 280 mm $\times$ 280 mm).

Only 1416 cases had all the information needed to be further analyzed. Based on the analyses of the data, subjects were divided into five main groups based on the answers of the questionnaires at the imaging time for classification purposes. MRI studies were perceived as symptomatic (clinically relevant pain) if the frequency $\geq$ 30 days and intensity $\geq 6/10$. If the frequency $\leq$ 7 days, intensity $\leq 3/10$, and no previous pain episodes in the follow-up period then it was set as asymptomatic. When there is pain at the moment and the intensity $\leq 5/10$ then MRI studies were perceived as slightly symptomatic (*N*=363). If the subject is asymptomatic at the moment of imaging but was symptomatic in the last 12 months then it was set as symptomatic/asymptomatic (*N*=527) (Table 3).

Two radiologists who were experienced in spine MRI performed grading on the lumbar disc independently. The radiologists reviewed each intervertebral disc from L1–L2 to L5-S1 by the Pfirrmann ([34]), and Modic [32] criteria (Table 4). To summarize, *N*=526 MRI studies with 113 symptomatic and 413 asymptomatic that contain all the corresponding data from MR images to Pfirrmann grades and Modic changes were used in this thesis (See table 4).

Table 3. Distribution statistics from the dataset.

| Groups | Distribution |
|---|---|
| **Asymptomatic** | **413 (29%)** |
| Slightly Symptomatic | 363 (25%) |
| **Symptomatic** | **113 (8%)** |
| Symptomatic/Asymptomatic | 527 (37%) |
| Not enough data for classification | 11 (1 %) |

Table 4. Pfirrmann changes and Modic changes statistics from the dataset used in this study.

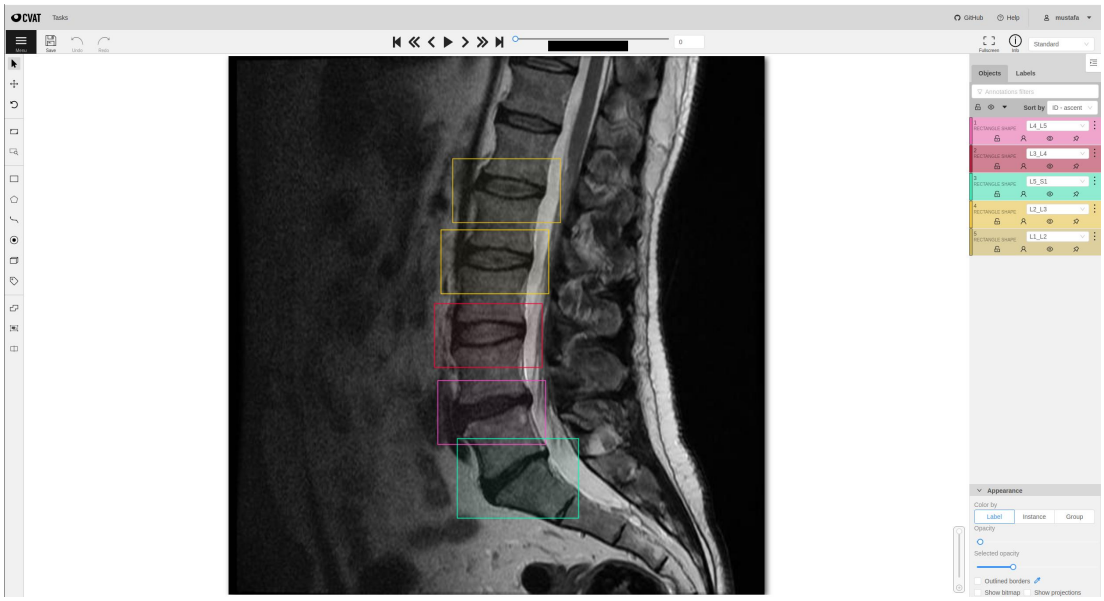
| Changes | L1-L2 | L2-L3 | L3-L4 | L4-L5 | L5-S1 |
|---|---|---|---|---|---|
| Pfirrmann 0 | 2 (< 1%) | 2 (< 1%) | 2 (< 1%) | 2 (< 1%) | 2 (< 1%) |
| Pfirrmann 1 | 0 (0%) | 0 (0%) | 1 (< 1%) | 1 (< 1%) | 0 (0%) |
| Pfirrmann 2 | 454(86%) | 412 (78%) | 332 (63%) | 180 (34%) | 151 (29%) |
| Pfirrmann 3 | 58 (11%) | 91(17%) | 131 (25%) | 161 (31%) | 126(24%) |
| Pfirrmann 4 | 11 (2%) | 20(4%) | 55 (10%) | 153 (29%) | 184(35%) |
| Pfirrmann 5 | 1 (< 1%) | 1 (< 1%) | 5 (1%) | 29 (6%) | 63(12%) |
| Modic 0 | 508 (97%) | 507 (96%) | 485 (92%) | 421 (80%) | 371 (71%) |
| Modic 1 | 7 (1%) | 9 (2%) | 14 (3%) | 20 (4%) | 31 (6%) |
| Modic 2 | 11 (2%) | 9 (2%) | 25 (5%) | 65 (12%) | 92 (17%) |
| Modic 3 | 0 (0%) | 1(< 1%) | 2 (< 1%) | 20 (4%) | 32 (6%) |

### 3.2. Data Annotation

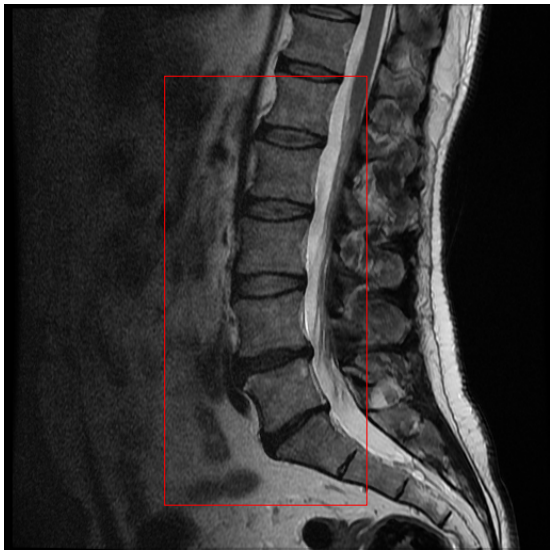
Some of the images may not have a clear visualization of the lumbar vertebrae. From the image stacks, the mid-sagittal slice (i.e. the location of the sagittal slice that went through the middle of the spinal canal) was first localized using a custom-made Python code. A script using python programming language was developed to extract the middle 2D image (mid-sagittal) slices and the images next to it ($\pm 2$). The selection of mid-sagittal images was done by choosing the most similar images to a couple of template middle-slice images that were manually chosen. The similarity was estimated by computing the correlation between the images using the fast Fourier transform method ([86]). The images are compared with a couple of template middle-slice images that were manually chosen. After verifying, the extracted location of mid-sagittal images, the location of five images had to be corrected.

After locating the mid-sagittal slice the area of interest i.e. lumbar region was extracted/cropped (Figure 3.14(b)). The unwanted regions were removed to help the model focus on the lumbar region only. In order to define the lumbar region, image cropping was performed. First, a subset of the data ($N$=331 mid-sagittal slices) was annotated manually using the CVAT-annotation tool ([87]) so that the subset annotated

dataset could be used to train segmentation CNN (Figure 3.14(a)). Only the mid-sagittal slices were annotated and used to train the segmentation pipeline. First, the independent annotation masks were made by drawing five rectangles covering the lumbar spine intervertebral discs starting from the middle of the L1-vertebral body to the middle of the L2-vertebral body, etc., until the middle of the S1-vertebral body (Figure 3.14(a)). Then the information from the five independent masks where combined so that the whole lumbar region was annotated (Figure 3.14(b)). This was done by finding the centroid of the independent masks and assigning a fixed size ROI of 200 $\times$ 400 to the center of the centroid point. The reason for the fixed size of that the input image size has to be the same as the neural network model.




(a)



(b)

Figure 14. CVAT annotation tool: (a) shows the user interface of the tool. The annotation was done by clicking on Create a rectangle for each class. (b) is the output images after processing the masks.

## 3.3. Segmentation of Cropped Images

### *3.3.1. Data Processing and Augmentation*

The image segmentation pipeline used the annotated subset ($N$=331) to train the segmentation DL model. Light data augmentations were employed to the training dataset to include variability. The augmentations were added to the training data only. Random horizontal flipping and random rotation between $-15$ and $15$ degrees were used.

### *3.3.2. Model Training and Validation Scheme*

The pipeline used U-Net architecture [80] where the encoder was an ImageNet [74] pre-trained (see Section 2.4.8) VGG11-BN [75] (Figure 13).

K-fold cross-validation ($K$=5) was implemented on the training data, see Figure 12. Out-of-fold (OOF) script was developed in order to average the predictions from each set of folds and average them to assess the performance on validation data.

The model was optimized by minimizing the binary cross-entropy loss function which is defined as

$$L_n = -(y \log(p) + (1 - y) \log(1 - p)), \tag{8}$$

where $y$ is the true label (0 or 1), and $p$ is the predicted probability. The performance evaluation was done using the dice coefficient. The value of the dice coefficient is defined as

$$DSC = \frac{2(x \cap y)}{x + y}, \tag{9}$$

where $y$ is the ground truth and $x$ is the prediction. The dice coefficient is calculating how many similar pixels in the prediction and the ground truth the algorithm finds and penalizes for the false positives.

After evaluating the model on the validation data, all the images were put together, and trained the final model to segment all of the remaining data ($N$=195). The segmentation pipeline output a mask to each middle image in each study in order to be used to segment the wanted (lumbar) region from spine images. Finally, the centroid pixel of the segmented masks was used as the center pixel to make $200 \times 400$ masks in order to have the same size masks.

## 3.4. Classification: Experimental Setup

### *3.4.1. Data Splitting*

The dataset was split to train ($N$=473, 90%) and test ($N$=53, 10%). During the split, both datasets (train and test) had the same percentage of symptomatic and asymptomatic cases, and images from the same MRI study would be in the same

dataset in the 3- and 5-slice approaches from the same study were used. The training dataset was split to train and validation using 10-fold cross-validation but one validation fold had significantly higher scores compared to the other folds. Thus, We ended up using only one fold as a validation fold. No data augmentations were employed.

The stratified group K-Fold ($K$=10) method was used to split the training-validation dataset during training. This split method was used to have the same percentage of symptomatic and asymptomatic cases at all folds, and also images from the same MRI study would be at the same fold in the models where multiple slices from the same study were used.

### *3.4.2. Experiments*

The automatically cropped ($200 \times 400$) images were used in the experiments. The images were standardized was done per image by using the mean and standard deviation from each image before analyses. The pipeline used ResNet50 (ImageNet pre-trained, see Section 2.4.8) model (Figure 15) and used the asymptomatic and symptomatic cases as input in order to binary classify $T_2$-weighted MRIs to symptomatic/asymptomatic cases. The model input consists of three-channel (colored) images similar to RGB images in order to use the pre-trained setting more efficiently, as the weights have been trained for a specific input configuration. Each mid-sagittal slice from the MRI studies was copied and stacked to make 3-channel images.

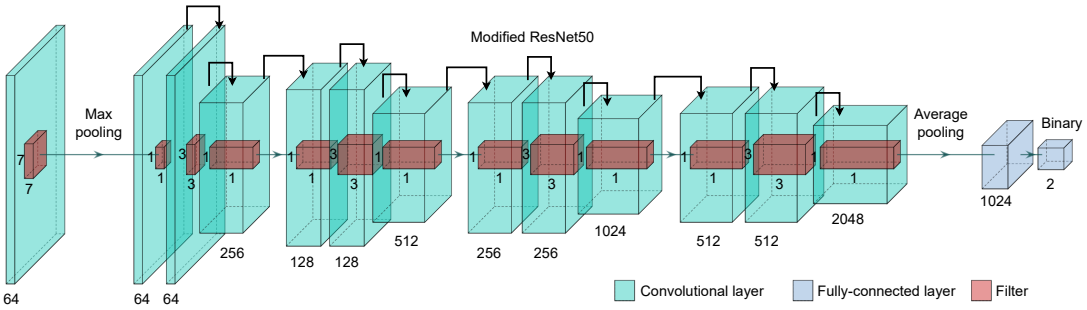



Figure 15. Network architecture: A diagram representation of a modified ResNet50 architecture. The numbers refer to the size and amount of kernels in each convolution layer.

Three approaches were investigated where the dataset was increased by adding the adjacent slices in order to showcase how many slices are needed to detect LBP from MR images:

- The lumbar spine region mid-sagittal slice images ($N$=526, 1-slice)
- The mid-sagittal slice images $\pm$ 1 were added to the dataset as an extra input ($N$=1 578, 3-slice)
- The mid-sagittal slice images $\pm$ 2 were added similar to earlier ($N$=2 630, 5-slice)

In the 3- and 5-slice approaches, individual MRI study was considered symptomatic if the DL model would predict pain in one of the slices.

The models were individually trained for 100 epochs in each approach but were stopped one epoch after unfreezing the weights reaching minimum loss as models would overfit if trained for more epochs. The weights were frozen before the fully-connected layer and run for 40 epochs before unfreezing in training. Binary cross-entropy loss was used (Equation 8). The model was optimized by updating network weights using Adam optimization algorithm [67].

The results from the validation and value difference of the probabilities (symptomatic and asymptomatic) made by the model were used to choose the hyperparameters used for the final model. The optimal threshold was computed based on the ROC curve. The ROC curve can give a graph summary of the model's skills at all classification thresholds. The validation dataset was used to compute the threshold that was used in the testing phase. This process was repeated five times by using different seed values each time to ensure generalization. The model with the best five Matthews Correlation Coefficient (MCC) results before the minimum loss epoch was saved each time.

The average results from the five models and thresholds were calculated to assess the performance using the test dataset. The performance of the DL models was compared using the average precision score (AP), ROC-AUC score, precision score, negative predictive value (NPV), recall, specificity (TNR), balanced accuracy score (BACC), and Matthews correlation coefficient (MCC) (Table 2). The losses for the 1-slice (Figure 16) 3-slice (Figure 17) and 5-slice (Figure 18) approaches have shown slight overfitting.

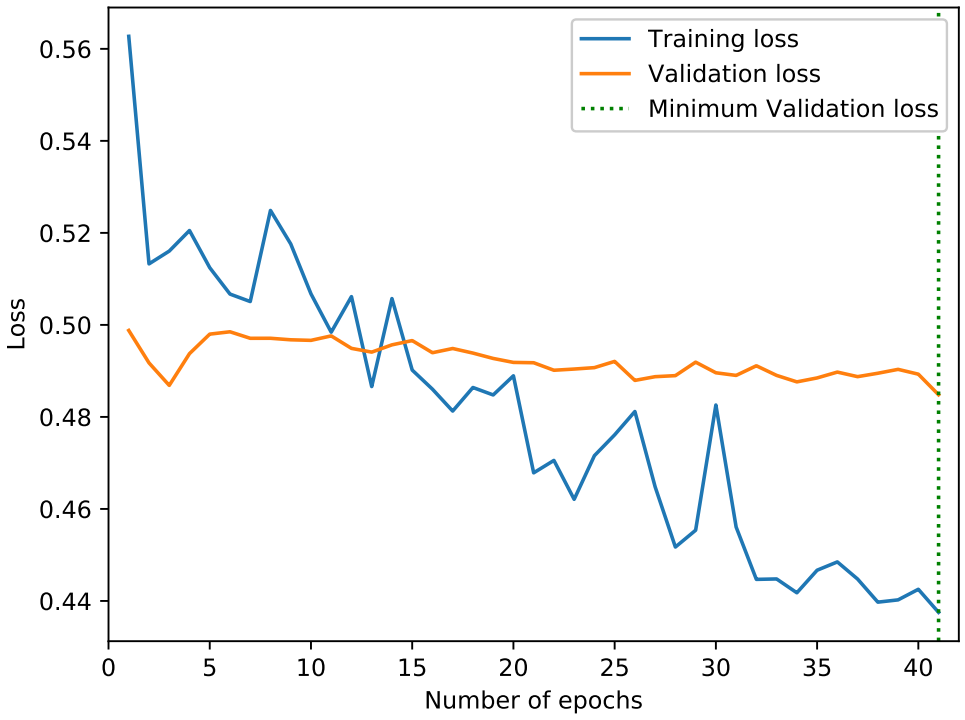



Figure 16. The learning curves for the mid-sagittal slice dataset approach.

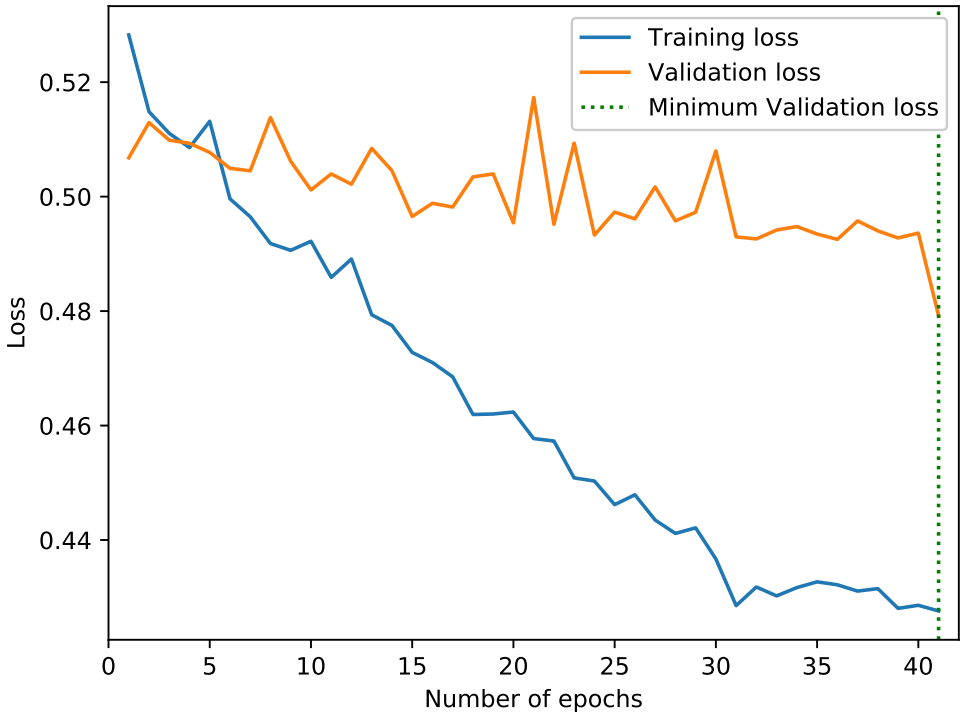



Figure 17. The learning curves for the mid-sagittal slice $\pm$ 1 dataset approach.

### 3.5. Deep Learning Framework and Hyperparameters

All the experiments were done using Pytorch machine learning library v.1.7.0 [70] in Python v.3.8 in order to build a DL-based classification pipeline.

For the segmentation pipeline, the number of epochs was 41, the learning rate was $1e{-}4$, weight decay was $1e{-}3$, the learning rate was reduced on the 20 epoch to $1e{-}5$ and on the 40 epoch to $1e{-}6$, test size was 10%, the number of folds was 5, batch size was 16, and the number of workers was 20.

For the classification pipeline learning rate was $1e{-}4$, weight decay was $1e{-}3$, on the 30 epoch the learning rate was reduced to $1e{-}5$, the number of epochs was 41 (40 epochs where the weights were frozen, and one epoch after unfreezing), test size was 10%, number of folds was 10, batch size was 32, and number of workers was 20.

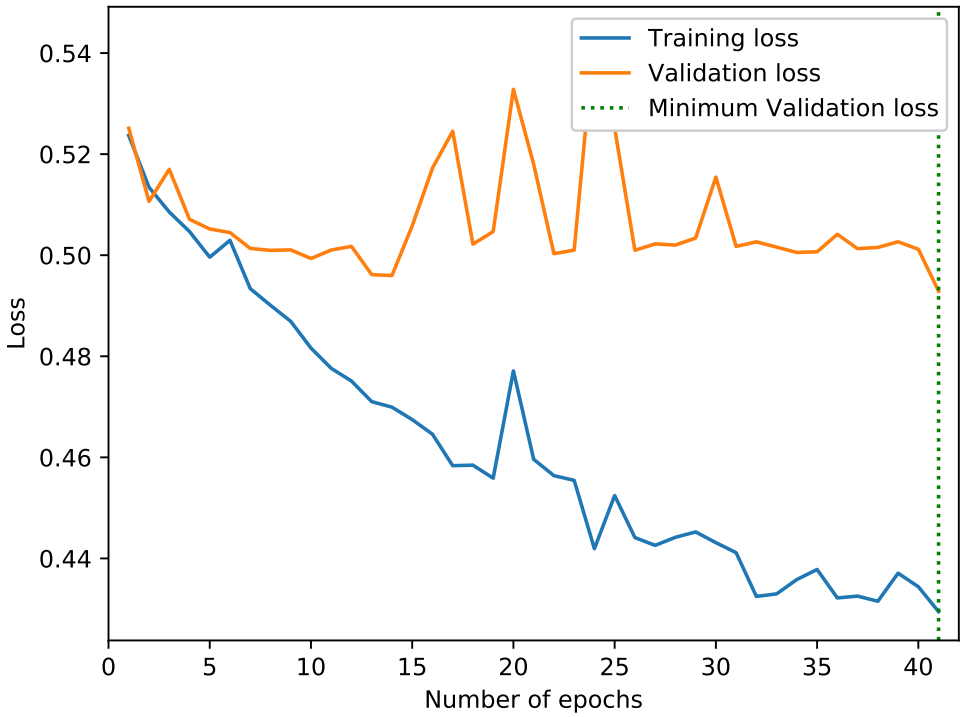



Figure 18. The learning curves for the mid-sagittal slice images $\pm$ 2 dataset approach.

# 4. RESULTS

## 4.1. Segmentation of Lumbar Region

By using 331 sagittal middle slice images to train the model, the Dice coefficient was calculated for the 5-folds, averaging 0.97 with a standard error of the mean of ± 0.01. After that, we used the pipeline to crop the rest (*N*=195) of the images (Figure 19).

## 4.2. Effect of Dataset Size in Low Back Pain Detection

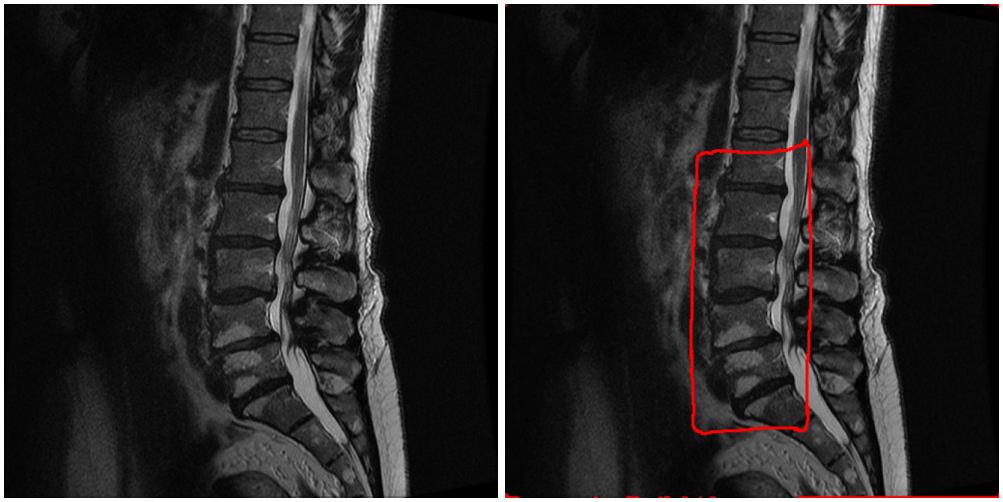
The dataset size has been experimented with as follows: The mid-sagittal slice from each MRI study (*N*=526, 1-slice), mid-sagittal slice images ±4 mm (*N*=1 578, 3-slice), and mid-sagittal slice images ±4 mm and ±8 mm (*N*=2 630, 5-slice). The dataset containing 3 adjacent slices per study had the best performance when assessed using different performance metrics (Table 5). The best dataset size was defined based on the metrics shown in table 5.

Table 5. The results of different dataset sizes using different metrics. AP stands for the average precision score, receiver operating characteristic area under curve (ROC-AUC) score, precision for the precision score, NPV for the negative predictive value score, Recall score, TNR for the specificity score, ACC for the balanced accuracy score, and MCC for the Matthews correlation coefficient.

| Metrics | 1-slice | 3-slice | 5-slice |
|---|---|---|---|
| AP | 0.403 ± 0.007 | 0.467 ± 0.025 | 0.369 ± 0.009 |
| ROC-AUC | 0.667 ± 0.008 | 0.740 ± 0.008 | 0.619 ± 0.011 |
| Precision | 0.500 ± 0.101 | 0.571 ± 0.025 | 0.393 ± 0.010 |
| NPV | 0.755 ± 0.019 | 0.846 ± 0.007 | 0.880 ± 0.011 |
| Recall | 0.143 ± 0.111 | 0.571 ± 0.036 | 0.786 ± 0.036 |
| TNR | 0.949 ± 0.050 | 0.846 ± 0.026 | 0.564 ± 0.038 |
| BACC | 0.546 ± 0.032 | 0.709 ± 0.011 | 0.675 ± 0.008 |
| MCC | 0.153 ± 0.073 | 0.418 ± 0.022 | 0.309 ± 0.014 |

In the 1-slice dataset, the model was able to predict most of the asymptomatic cases, however, it also misclassified most of the symptomatic cases (Figure 20(a)). The use of the sagittal middle slice images only does not provide enough information to the model in order to classify LBP. In the 5-slice dataset, the model was able to predict most of the symptomatic cases, but it also misclassified many of the asymptomatic cases (Figure 22(a)).

In the 1-slice approach, the ROC curve indicated that the model is able to predict true negatives but has low true positives across different thresholds (Figure 20(b)). For the 3-slice approach, the ROC-AUC curve had the highest ROC-AUC score indicating that the model is able to predict more true negatives than false negatives and slightly higher true positives (Figure 21(b)). Where for the 5-slice approach, the ROC-AUC curve

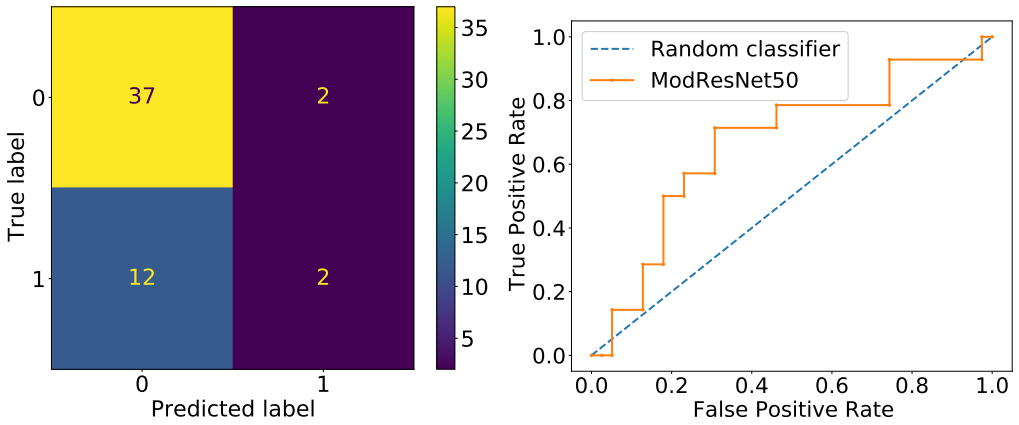


(a) (b)

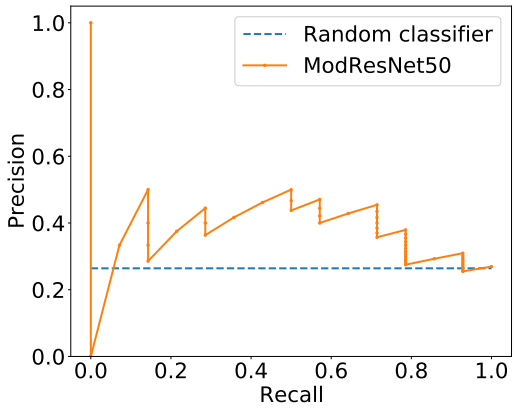
Figure 19. Example of a (b) segmentation output before processing to a fixed mask size of $200 \times 400$ from (a) the input image.

indicates that the model predicts low false positives and slightly higher true negatives across different thresholds (Figure 22(b)).

The precision-Recall curve was also computed to assess how different models are good at predicting the positive class. The 3-slice curve (Figure 21(c)) is the best compared to the 1-slice curve (Figure 20(c)) and 5-slice curve (Figure 22(c)), indicating that the model is able to predict the positive class for different thresholds.




(a) (b)




(c)

Figure 20. The results of the 1-slice dataset where (a) is the confusion matrix (b) is the ROC-AUC curve (ROC-AUC score = 0.667 $\pm$ 0.008), and (c) is the Precision-Recall Curve (AP score = 0.403 $\pm$ 0.007). Confusion matrix of the 1-slice approach, where 0 stands for the negative (asymptomatic) cases and 1 for positives (symptomatic) cases. The model correctly predicts 2 symptomatic and 37 asymptomatic cases and incorrectly predicts 12 symptomatic and 2 asymptomatic cases. The dotted line in the plot related to the Precision-Recall curve shows the proportion of positive samples in the dataset.

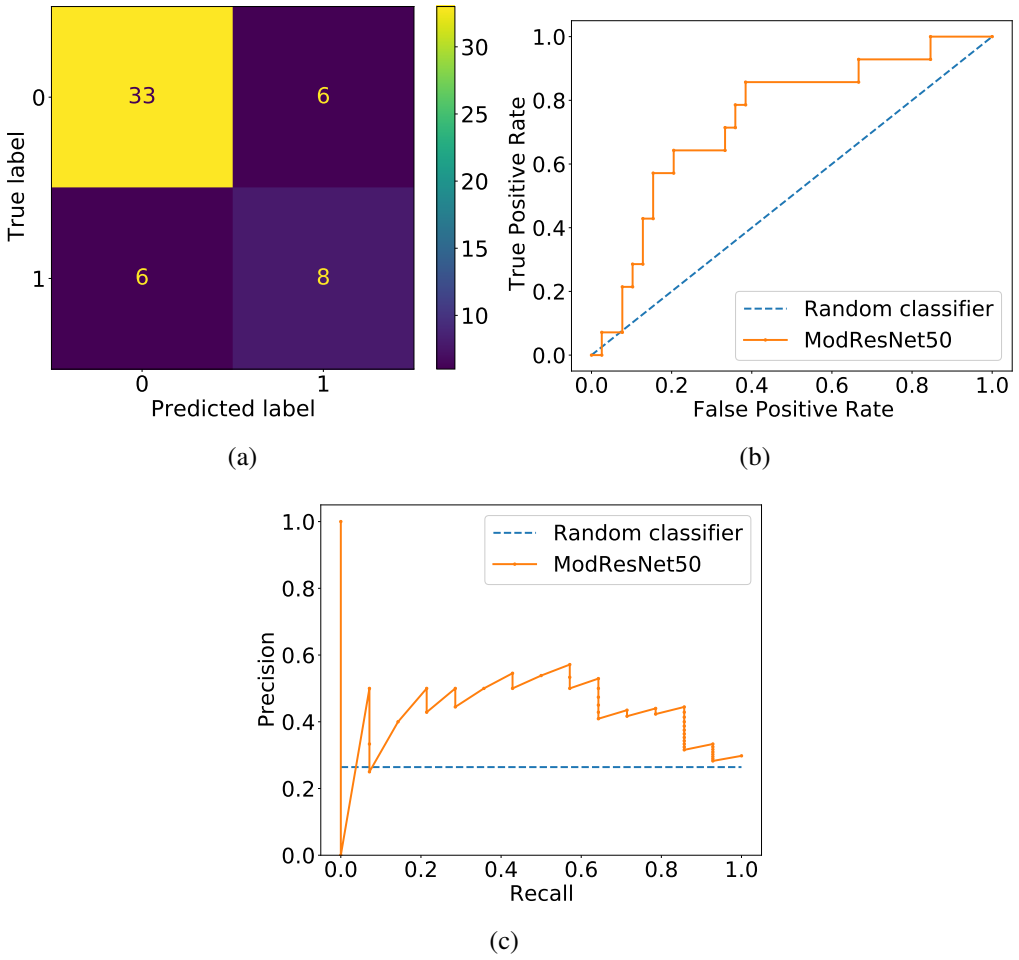
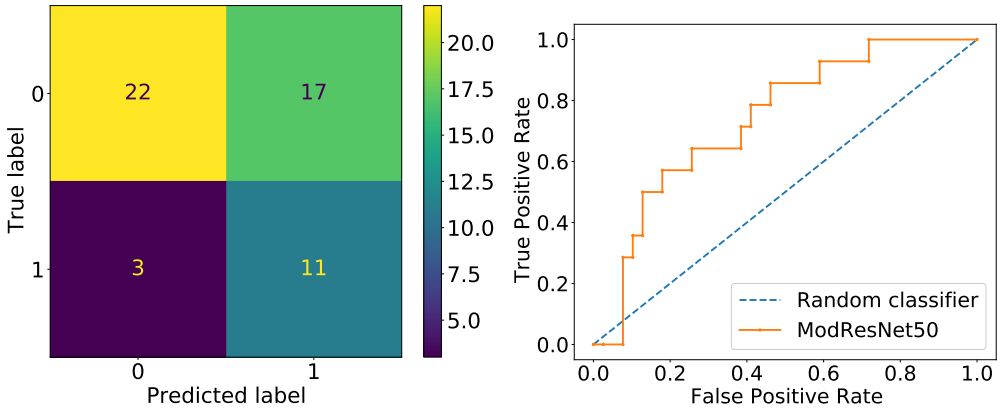
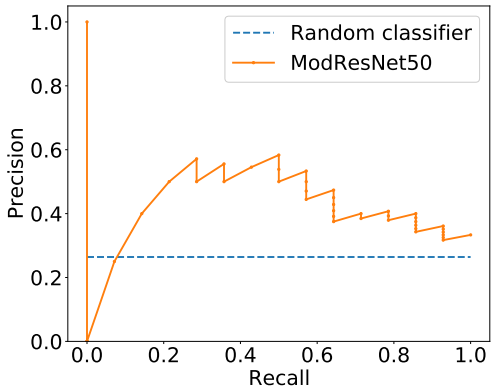



Figure 21. The results of the 3-slice dataset where (a) is the confusion matrix (b) is the ROC-AUC curve (ROC-AUC score = 0.724 $\pm$ 0.005), and (c) is the Precision-Recall Curve (AP score = 0.469 $\pm$ 0.017). The dotted line in the plot related to the Precision-Recall curve shows the proportion of positive samples in the dataset.




(a) (b)




(c)

Figure 22. The results of the 5-slice dataset where (a) is the confusion matrix (b) is the ROC-AUC curve (ROC-AUC score = 0.619 $\pm$ 0.004), and (c) is the Precision-Recall Curve (AP score = 0.369 $\pm$ 0.004). The dotted line in the plot related to the Precision-Recall curve shows the proportion of positive samples in the dataset.

# 5. DISCUSSION

## 5.1. Main Findings

In this work, DL CNN model was used to investigate how it can be applied to LBP classification from $T_2$-weighted MRI studies of symptomatic and asymptomatic cases. Detection of LBP from MR images can be a very challenging task for medical professionals because the feeling of pain is subjective and not necessarily related to any pathological change. The hypothesis was to investigate whether DL can be used to detect lower back pain from the lumbar MRI of symptomatic and asymptomatic cohort subjects as DL enables feature learning from MRI textures. Before classifying the images, a U-Net [80] network with VGG11-BN [75] encoder was used to crop the images to remove unwanted regions around the lumbar spine. The acquired results showcase that DL-based LBP classification is capable of identifying symptomatic cases from the NFBC1966 data better than a random classifier (See figures 20 21 22). We found that the 3-slices approach achieved the best results as the amount of data could be increased compared to only the mid-sagittal slice. The three slices could be sufficiently close to the anatomical regions of the spine where there are degenerative changes related to LBP. The baseline model using only the mid-sagittal slice from each MRI study (1 slice) for classification achieved the lowest classification scores. The low classification scores may be because of the small viewpoint as changes related to LBP may not be visible in the mid-sagittal slice. Adding more data in by adding even more slices (5-slice) seemed to worsen the accuracy of the model, as more false positives were detected (Figure 22). The worsened accuracy results may be because of the large amount of data added that has no MRI findings nor degenerative changes related to LBP. The losses for the 1- (Figure 16), 3- (Figure 17), and 5-slice (Figure 18) approaches have demonstrated overfitting due to the used hyperparameter yet provided better results than other cases where loss curves did not demonstrate overfitting. The reason may be that the model learned the incorrect features for LBP as the curve may show that the model is learning, but does not disclose if the model is learning the wanted features.

The model trained using data from the 3-slices approach was able to correctly predict more than half of the symptomatic and asymptomatic cases. There are also relatively many misclassified cases. Sometimes in diagnostics, models predicting false positives may be considered better than models predicting false negatives (at least in this case the disease will not go unnoticed), although neither are wanted and in some cases. In the case of low back pain diagnostics, it is known that a lot of false positives are found when only degenerative changes in the back are considered, making the optimal trade-off complicated to deduce. Therefore, low back pain diagnostics could benefit from a DL-based tool.

To conclude, feature analysis of lumbar spine MRI data using DL shows promise as a diagnostic tool in the assessment of LBP. This methodology could be used, for example, to identify which tissues and anatomical regions account for the presence of LBP, or to process clearly negative cases to lighten the diagnostic workflow of medical professionals in routine imaging tasks. The continuous investigation of LBP and the automated methods to find and localize it has the potential to improve its management, and at the same time help radiologists in the diagnosing and treating processes.

Several methods were experimented with to improve the model performance but did not help thus they were excluded from the study. An example of such (discarded) technique was the synthetic minority oversampling technique (SMOTE) technique, where augmented copies from the original images are added to increase the minority class (symptomatic). Many augmentations augmentation techniques (like; random noise, random rotation with an angle range of $(-10, 0)$, pad between $5 - 10$ pixels and randomly crop to the right size, apply contrast limited adaptive histogram equalization (CLAHE) to the input image and blur the image slightly) were tested both individually and together where augmentation was added randomly but did not show any remarkable improvement in model performance. In addition, many other CNN architectures such as VGG11-BN [75], RESNET34 [64], RESNET101 [64], etc, were tested but did not improve the results and thus were disqualified. The tested methods did not improve model prediction, and that may be because of the over-complexity of the models. In this problem, simple and straightforward methods performed better than more complicated ones.

Gradient-weighted Class Activation Mapping (Grad-CAM), is one of the visual explanations methods used in this study to show where CNN models focus the most when giving a prediction [88]. The heatmaps can give an understanding of how the model is training. Also, they can give feedback on why the model had trouble making the correct classification in other cases. The heatmaps were logical in some cases and not logical in others making them not useful in clinical applications and thus, were excluded from the study.

## 5.2. Related Work

Compared to a previous study where texture analysis-based machine learning was utilized to classify LBP using the same NFBC1966 data [10], the results in this work showcased lower scores except for negligibly higher precision in the 3-slice model. Yet, in the present study, a larger region of interest was used that includes the entire lumbar spine with some surrounding tissues rather than segmenting the vertebrae and IVDs. Moreover, in the presented study a single convolutional neural network that could automatically analyze lumbar MRI data compared with the data processing pipeline (segmentation - texture analysis - principal component analysis - logistic regression classification).

Another study investigated the use of a neural network classifier on kinematic data to develop a LBP classification framework [58]. The classifier was able to utilize neural networks to predict LBP with $85\%$ accuracy. Finally, another study presented the effectiveness of neural networks in categorizing patients with acute and chronic LBP based on activity meter data and electronic symptom diary [89]. However, in the present work, convolutional neural networks were used to detect LBP from MR images directly without the need for any other test. The use of MR images may provide more information on the cause of pain, also MRI is the most used method in the search for LBP.

### 5.3. Limitations and Future Work

LBP is subjective thus, finding it from images automatically is quite challenging. The small dataset available for this study made the task of finding LBP challenging for the DL model. The imbalanced dataset (Symptomatic *N*=113, Asymptomatic *N*=413) made it challenging to learn the features of the symptomatic cases as the model did not have many examples of symptomatic cases.

This study is not by all means complete in the sense that more future work is needed to improve it even further. We aim to build on this work to overcome its limitations. Specifically, exploring the possibility of using three-dimensional images as input to the model, where each image input would have 3, 5, or up to 9 adjacent slices. Furthermore, we aim to find the optimal data augmentations to improve the model's ability to detect symptomatic and asymptomatic cases. We aim to build a specific neural network to detect LBP and pre-trained it using similar data (spine MRI). In addition, we aim to experiment using more input data such as Modic and Pfirrmann changes and their effect on improving model performance. We aim to investigate the use of different MRI contrast for example $T_1$- and $T_2$-weighted to improve model performance. Finally, we aim to showcase the location of the specific features that are most correlated with LBP by implementing a heatmap-style visualization for maximal activations in the final layers.

# 6. SUMMARY

To conclude, in this study the lumbar spine region was investigated in order to detect LBP by analyzing $T_2$-weighted MRI using CNN-based methods. The Northern Finland Birth Cohort 1966 dataset was used as the input data for deep learning.

Image segmentation pipeline using handmade masks as inputs was used to automatically crop the lumbar spine region from the images. ResNet50 deep learning model was used to analyze the cropped MRI studies. Three dataset sizes were analyzed: The mid-sagittal slice from each MRI study with 526 images (1-slice), mid-sagittal slices and its immediate neighboring slices $\pm$4 mm with 1 578 images (3-slice), and five middle-most sagittal slices $\pm$4 mm and $\pm$8 mm with 2 630 images (5-slice).

The results showed that the model trained with the 3-slice dataset had the best performance. This provides a baseline for future work to use only three middle-most sagittal slices in the search for LBP. The analysis shows optimistic results that can be further developed to be used as a tool for diagnosing and assessing LBP.

Furthermore, this study suggests that deep learning-based models are powerful tools and could be used for LBP diagnostics of lumbar spine MRI. The developed classification pipeline could be beneficial, for example, in pinpointing negative cases in order to improve the workflow of routine diagnostic imaging tasks.

# 7. REFERENCES


[1] Hoy D., Brooks P., Blyth F. & Buchbinder R. (2010) The epidemiology of low back pain. Best Practice and Research Clinical Rheumatology 24, pp. 769–781. URL: https://www.sciencedirect.com/science/article/pii/S1521694210000884, DOI: https://doi.org/10.1016/j.berh.2010.10.002.

[2] Andersson G.B.J. (1998) Epidemiology of low back pain. Acta Orthopaedica Scandinavica 69, pp. 28–31. URL: https://doi.org/10.1080/17453674.1998.11744790, DOI: 10.1080/17453674.1998.11744790.

[3] National Institutes of Health (2020), Low back pain fact sheet. URL: https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Low-Back-Pain-Fact-Sheet, Accessed: 18.6.2022.

[4] Sheehan N. (2010) Magnetic resonance imaging for low back pain: Indications and limitations. Postgraduate medical journal 86, pp. 374–8. DOI: 10.1136/ard.2009.110973.

[5] Patel N.D., Broderick D.F., Burns J., Deshmukh T.K., Fries I.B., Harvey H.B., Holly L., Hunt C.H., Jagadeesan B.D., Kennedy T.A., O'Toole J.E., Perlmutter J.S., Policeni B., Rosenow J.M., Schroeder J.W., Whitehead M.T., Cornelius R.S. & Corey A.S. (2016) Acr appropriateness criteria low back pain. Journal of the American College of Radiology 13, pp. 1069–1078. URL: https://www.sciencedirect.com/science/article/pii/S1546144016304409, DOI: https://doi.org/10.1016/j.jacr.2016.06.008.

[6] Cousins J.P. & Haughton V.M. (2009) Magnetic resonance imaging of the spine. JAAOS - Journal of the American Academy of Orthopaedic Surgeons 17. URL: https://journals.lww.com/jaaos/Fulltext/2009/01000/Magnetic_Resonance_Imaging_of_the_Spine.4.aspx.

[7] LeCun Y., Bottou L., Bengio Y. & Haffner P. (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, p. 2278 – 2323. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0032203257&doi=10.1109%2f5.726791&partnerID=40&md5=bcd0f8af84203ead7e96411039389e6b, DOI: 10.1109/5.726791.

[8] Jang H.J. & Cho K.O. (2019) Applications of deep learning for the analysis of medical data. Archives of pharmacal research 42, p. 492—504. URL: https://doi.org/10.1007/s12272-019-01162-9, DOI: 10.1007/s12272-019-01162-9.

[9] Jamaludin A., Kadir T. & Zisserman A. (2017) Spinenet: Automated classification and evidence visualization in spinal MRIs. Medical Image Analysis

41, pp. 63–73. URL: https://www.sciencedirect.com/science/article/pii/S136184151730110X, DOI: https://doi.org/10.1016/j.media.2017.07.002.

[10] Ketola J.H.J., Inkinen S.I., Karppinen J., Niinimäki J., Tervonen O. & Nieminen M.T. (2021) T2-weighted magnetic resonance imaging texture as predictor of low back pain: A texture analysis-based classification pipeline to symptomatic and asymptomatic cases. Journal of Orthopaedic Research n/a. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jor.24973, DOI: https://doi.org/10.1002/jor.24973.

[11] Jarvik J., Hollingworth W., Heagerty P., Haynor D. & Deyo R. (2001) The longitudinal assessment of imaging and disability of the back (LAIDBack) study. Spine 26, pp. 1158–66. DOI: 10.1097/00007632-200105150-00014.

[12] Kjaer P., Bendix T., Sorensen J., Korsholm L. & Leboeuf-Yde C. (2007) Are MRI-defined fat infiltrations in the multifidus muscles associated with low back pain? BMC medicine 5, p. 2. DOI: 10.1186/1741-7015-5-2.

[13] Lynch P.J. (2020), File:segments of vertebrae.svg. URL: https://commons.wikimedia.org/wiki/File:Segments_of_Vertebrae.svg, Accessed: 2.8.2022.

[14] Kandahari A.M., Puvanesarajah V., Shen F.H., Raso J. & Hassanzadeh H. (2022) 1 - anatomy of the spine. In: D. Samartzis, J.I. Karppinen & F.M. Williams (eds.) Spine Phenotypes, Academic Press, pp. 1–34. URL: https://www.sciencedirect.com/science/article/pii/B9780128227787000055, DOI: https://doi.org/10.1016/B978-0-12-822778-7.00005-5.

[15] Jmarchn (2015), File:vertebra superior view-en.svg. URL: https://commons.wikimedia.org/wiki/File:Vertebra_Superior_View-en.svg, Accessed: 6.8.2022.

[16] Jmarchn (2015), File:vertebra posterolateral-en.svg. URL: https://commons.wikimedia.org/wiki/File:Vertebra_Posterolateral-en.svg, Accessed: 3.8.2022.

[17] Malik K. & Nelson A. (2018) Chapter 24 - overview of low back pain disorders. In: H.T. Benzon, S.N. Raja, S.S. Liu, S.M. Fishman & S.P. Cohen (eds.) Essentials of Pain Medicine (Fourth Edition), Elsevier, fourth edition ed., pp. 193–206.e2. URL: https://www.sciencedirect.com/science/article/pii/B9780323401968000243, DOI: https://doi.org/10.1016/B978-0-323-40196-8.00024-3.

[18] Braun J., Baraliakos X., Regel A. & Kiltz U. (2014) Assessment of spinal pain. Best Practice & Research Clinical Rheumatology 28, pp. 875–887. URL: https://www.sciencedirect.com/science/article/pii/S1521694215000388, DOI: https://doi.org/10.1016/j.berh.2015.04.031.

[19] Luoma K., Riihimäki H., Luukkonen R., Raininko R., Viikari-Juntura E. & Lamminen A. (2000) Low back pain in relation to lumbar disc degeneration. Spine 25, pp. 487–92. DOI: 10.1097/00007632-200002150-00016.

[20] Allegri M., Montella S., Salici F., Valente A., Marchesini M., Compagnone C., Baciarello M., Manferdini M. & Fanelli G. (2016) Mechanisms of low back pain: a guide for diagnosis and therapy. F1000Research 5, p. 1530. DOI: 10.12688/f1000research.8105.1.

[21] Koes B., Tulder M. & Thomas S. (2006) Diagnosis and treatment of low back pain. BMJ (Clinical research ed.) 332, pp. 1430–4. DOI: 10.1136/bmj.332.7555.1430.

[22] Yang H., Liu H., min Li Z., Zhang K., Wang J., Wang H. & Zheng Z. (2015) Low back pain associated with lumbar disc herniation: role of moderately degenerative disc and annulus fibrous tears. International journal of clinical and experimental medicine 8 2, pp. 1634–44. URL: https://pubmed.ncbi.nlm.nih.gov/25932092/.

[23] Goode A., Carey T. & Jordan J. (2013) Low back pain and lumbar spine osteoarthritis: How are they related? Current rheumatology reports 15, p. 305. DOI: 10.1007/s11926-012-0305-z.

[24] Wong A.Y.A.Y., Karppinen J.J. & Samartzis D.D. (2017) Low back pain in older adults: risk factors, management options and future directions, vol. 12. Scoliosis and Spinal Disorders. URL: http://urn.fi/urn:nbn:fi-fe201706277485.

[25] Videman T., Battie M., Gibbons L., Maravilla K., Manninen H. & Kaprio J. (2003) Associations between back pain history and lumbar MRI findings. Spine 28, pp. 582–588. DOI: 10.1097/01.BRS.0000049905.44466.73.

[26] Dagenais S., Caro J. & Haldeman S. (2008) A systematic review of low back pain cost of illness studies in the united states and internationally. The Spine Journal 8, pp. 8–20. URL: https://www.sciencedirect.com/science/article/pii/S1529943007008984.

[27] Guo H.R., Tanaka S., Halperin W.E. & Cameron L.L. (1999) Back pain prevalence in us industry and estimates of lost workdays. American Journal of Public Health 89, pp. 1029–1035. URL: https://doi.org/10.2105/AJPH.89.7.1029.

[28] Katz J. (2006) Lumbar disc disorders and low-back pain: Socioeconomic factors and consequences. The Journal of bone and joint surgery. American volume 88 Suppl 2, pp. 21–4. DOI: 10.2106/JBJS.E.01273.

[29] Dieleman J.L., Baral R., Birger M., Bui A.L., Bulchis A., Chapin A., Hamavid H., Horst C., Johnson E.K., Joseph J., Lavado R., Lomsadze L., Reynolds A., Squires E., Campbell M., DeCenso B., Dicker D., Flaxman A.D., Gabert R., Highfill T., Naghavi M., Nightingale N., Templin T., Tobias M.I., Vos T. & Murray C.J.L.

(2016) US Spending on Personal Health Care and Public Health, 1996-2013. JAMA 316, pp. 2627–2646. URL: https://doi.org/10.1001/jama.2016.16885.

[30] Olafsson G., Jonsson E., Fritzell P., Hägg O. & Borgström F. (2018) Cost of low back pain: results from a national register study in sweden. European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society 27, p. 2875—2881. URL: https://doi.org/10.1007/s00586-018-5742-6, DOI: 10.1007/s00586-018-5742-6.

[31] Bajwa Z.H., Gupta S., Warfield C.A. & Steinman T.I. (2001) Pain management in polycystic kidney disease. Kidney International 60, pp. 1631–1644. URL: https://www.sciencedirect.com/science/article/pii/S0085253815480433.

[32] Modic M.T., Steinberg P.M., Ross J.S., Masaryk T.J. & Carter J.R. (1988) Degenerative disk disease: assessment of changes in vertebral body marrow with mr imaging. Radiology 166, pp. 193–199. URL: https://doi.org/10.1148/radiology.166.1.3336678.

[33] Kuisma M., Karppinen J., Niinimäki J., Ojala R., Haapea M., Heliövaara M., Korpelainen R., Taimela S., Natri A. & Tervonen O. (2007) Modic changes in endplates of lumbar vertebral bodies: Prevalence and association with low back and sciatic pain among middle-aged male workers. Spine 32, pp. 1116–22. DOI: 10.1097/01.brs.0000261561.12944.ff.

[34] Pfirrmann C., Metzdorf A., Zanetti M., Hodler J. & Boos N. (2001) Magnetic resonance classification of lumbar intervertebral disc degeneration. Spine 26, pp. 1873–8. DOI: 10.1097/00007632-200109010-00011.

[35] Gordon R. & Bloxham S. (2016) A systematic review of the effects of exercise and physical activity on non-specific chronic low back pain. Healthcare 4, p. 22. DOI: 10.3390/healthcare4020022.

[36] Sprouse-Blum A.S., Smith G., Sugai D.Y. & Parsa F.D. (2010) Understanding endorphins and their importance in pain management. Hawaii medical journal 69 3, pp. 70–1. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3104618/.

[37] Carassiti M., Pascarella G., Strumia A., Russo F., Papalia G., Cataldo R., Gargano F., Costa F., Pierri M., De Tommasi F., Massaroni C., Schena E. & Agrò F. (2021) Epidural steroid injections for low back pain: A narrative review. International Journal of Environmental Research and Public Health 19, p. 231. DOI: 10.3390/ijerph19010231.

[38] Lumbar spinal fusion surgery. https://www.spine-health.com/treatment/spinal-fusion/lumbar-spinal-fusion-surgery. Accessed: 23.10.2022.

[39] Ibrahim T., Tleyjeh I.M. & Gabbar O.A. (2006) Surgical versus non-surgical treatment of chronic low back pain: a meta-analysis of randomised trials. International Orthopaedics 32, pp. 107–113. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2219937/.

[40] Leary S.P., Regan J.J., Lanman T.H. & Wagner W.H. (2007) Revision and explantation strategies involving the charitÉ lumbar artificial disc replacement. Spine 32, pp. 1001–1011. DOI: 10.1097/01.brs.0000260794.73938.93.

[41] Balagué F., Mannion A.F., Pellisé F. & Cedraschi C. (2012) Non-specific low back pain. The Lancet 379, pp. 482–491. URL: https://www.sciencedirect.com/science/article/pii/S0140673611606107.

[42] Westbrook C. & Talbot J. (2018) MRI in Practice. Wiley Publishing, 5th ed.

[43] McRobbie D.W., Moore E.A., Graves M.J. & Prince M.R. (2006) MRI from Picture to Proton. Cambridge University Press, 2 ed.

[44] Lauterbur P.C. (1973) Image formation by induced local interactions: Examples employing nuclear magnetic resonance. Nature 242, pp. 190–191. DOI: https://doi.org/10.1038/242190a0.

[45] Maravilla K.R., Lesh P., Weinreb J.C., Selby D.K. & Mooney V. (1985) Magnetic resonance imaging of the lumbar spine with ct correlation. American Journal of Neuroradiology 6, pp. 237–245. URL: http://www.ajnr.org/content/6/2/237.

[46] Beattie P. (2008) Current understanding of lumbar intervertebral disc degeneration: A review with emphasis upon etiology, pathophysiology, and lumbar magnetic resonance imaging findings. Journal of Orthopaedic & Sports Physical Therapy 38, pp. 329–340. URL: https://doi.org/10.2519/jospt.2008.2768.

[47] Takatalo J., Karppinen J., Niinimaki J., Taimela S., Nayha S., Mutanen P., Sequeiros R., Kyllonen E. & Tervonen O. (2011) Does lumbar disc degeneration on MRI associate with low back symptom severity in young finnish adults? Spine 36, pp. 2180–2189. DOI: 10.1097/BRS.0b013e3182077122.

[48] Määttä J., Wadge S., MacGregor A., Karppinen J. & Williams F. (2016) Vertebral endplate (modic) change is an independent risk factor for episodes of severe and disabling low back pain. Orthopaedic Proceedings 98-B, pp. 17–17. URL: https://online.boneandjoint.org.uk/doi/abs/10.1302/1358-992X.98BSUPP_6.SBPR2015-017, DOI: 10.1302/1358-992X.98BSUPP_6.SBPR2015-017.

[49] Koes B.W., van Tulder M.W. & Peul W.C. (2007) Diagnosis and treatment of sciatica. BMJ 334, pp. 1313–1317. URL: https://www.bmj.com/content/334/7607/1313.

[50] Goodfellow I., Bengio Y. & Courville A. (2016) Deep Learning. MIT Press. URL: http://www.deeplearningbook.org.

[51] Ivakhnenko A.G. & Lapa V.G. (1966), Cybernetic predicting devices. URL: https://gwern.net/docs/ai/1966-ivakhnenko.pdf.

[52] Wang S., Yang D.M., Rong R., Zhan X. & Xiao G. (2019) Pathology image analysis using segmentation deep learning algorithms. The American Journal of Pathology 189, pp. 1686 – 1698. URL: http://www.sciencedirect.com/science/article/pii/S0002944018311210.

[53] de Bruijne M. (2016) Machine learning approaches in medical image analysis: From detection to diagnosis. Medical Image Analysis 33, pp. 94–97. URL: https://www.sciencedirect.com/science/article/pii/S1361841516301098.

[54] Altaf F., Islam S.M.S., Akhtar N. & Janjua N.K. (2019) Going deep in medical image analysis: Concepts, methods, challenges, and future directions. IEEE Access 7, pp. 99540–99572. DOI: 10.1109/ACCESS.2019.2929365.

[55] Anwar S., Majid M., Qayyum A., Awais M., Alnowami M. & Khan K. (2018) Medical image analysis using convolutional neural networks: A review. Journal of Medical Systems 42, p. 226. DOI: 10.1007/s10916-018-1088-1.

[56] Kaur H., Pannu H.S. & Malhi A.K. (2019) A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM Comput. Surv. 52. URL: https://doi.org/10.1145/3343440.

[57] Parisa A., Taravat Y., C. B.E., Nayeb A.H., Shirzad A., Sohrab S. & Ali M. (2020) A review on the use of artificial intelligence in spinal diseases. Asian Spine J 14, pp. 543–571. URL: http://www.asianspinejournal.org/journal/view.php?number=1194, DOI: 10.31616/asj.2020.0147.

[58] Bishop J., Szpalski M., Ananthraman S., McIntyre D. & Pope M. (1997) Classification of low back pain from dynamic motion characteristics using an artificial neural network. Spine 22, p. 2991—2998. URL: https://doi.org/10.1097/00007632-199712150-00024.

[59] Albawi S., Mohammed T.A. & Al-Zawi S. (2017) Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET), pp. 1–6. DOI: 10.1109/ICEngTechnol.2017.8308186.

[60] Linnainmaa S. (1976) Taylor expansion of the accumulated rounding error. BIT 16, p. 146–160. URL: https://doi.org/10.1007/BF01931367.

[61] Schmidhuber J. (2015) Deep learning in neural networks: An overview. Neural Networks 61, pp. 85–117. URL: https://doi.org/10.1016%2Fj.neunet.2014.09.003.

[62] Anaya-Isaza A., Mera-Jiménez L. & Zequera-Diaz M. (2021) An overview of deep learning in medical imaging. Informatics in Medicine Unlocked 26, p. 100723. URL: https://www.sciencedirect.com/science/article/pii/S2352914821002033.

[63] Krizhevsky A., Ilya S. & E H.G. (2012) Imagenet classification with deep convolutional neural networks. In: F. Pereira, C.J.C. Burges, L. Bottou & K.Q. Weinberger (eds.) Advances in Neural Information Processing Systems 25, Curran Associates, Inc., pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[64] He K., Zhang X., Ren S. & Sun J. (2015), Deep residual learning for image recognition. URL: https://arxiv.org/abs/1512.03385.

[65] Bishop C.M. (2007) Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, 1 ed. URL: https://link.springer.com/book/9780387310732.

[66] Elson J., Douceur J.J., Howell J. & Saul J. (2007) Asirra: A captcha that exploits interest-aligned manual image categorization. In: Proceedings of 14th ACM Conference on Computer and Communications Security (CCS), Association for Computing Machinery, Inc., pp. 366–374. URL: https://www.microsoft.com/en-us/research/publication/asirra-a-captcha-that-exploits-interest-aligned-manual-image-categorization/.

[67] Kingma D.P. & Ba J. (2014), Adam: A method for stochastic optimization. URL: https://arxiv.org/abs/1412.6980.

[68] Robbins H.E. (1951) A stochastic approximation method. Annals of Mathematical Statistics 22, pp. 400–407. URL: https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-22/issue-3/A-Stochastic-Approximation-Method/10.1214/aoms/1177729586.full.

[69] Cox D.R. (1958) The regression analysis of binary sequences. Journal of the Royal Statistical Society. Series B (Methodological) 20, pp. 215–242. URL: http://www.jstor.org/stable/2983890.

[70] Paszke A., Gross S., Massa F., Lerer A., Bradbury J., Chanan G., Killeen T., Lin Z., Gimelshein N., Antiga L., Desmaison A., Kopf A., Yang E., DeVito Z., Raison M., Tejani A., Chilamkurthy S., Steiner B., Fang L., Bai J. & Chintala S. (2019) Pytorch: An imperative style, high-performance deep learning library. In: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox & R. Garnett (eds.) Advances in Neural Information Processing Systems 32, Curran Associates, Inc., pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[71] Zhang A., Lipton Z.C., Li M. & Smola A.J. (2020) Dive into Deep Learning. d2l.ai. https://d2l.ai.

[72] Srivastava N., Hinton G., Krizhevsky A., Sutskever I. & Salakhutdinov R. (2014) Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15, pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[73] Shorten C. & Khoshgoftaar T.M. (2019) A survey on image data augmentation for deep learning. Journal of Big Data 6, pp. 1–48. DOI: 10.1186/S40537-019-0197-0.

[74] Deng J., Dong W., Socher R., Li L.J., Li K. & Fei-Fei L. (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, Ieee, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[75] Simonyan K. & Zisserman A. (2014), Very deep convolutional networks for large-scale image recognition. URL: https://arxiv.org/abs/1409.1556, DOI: 10.48550/ARXIV.1409.1556.

[76] Krizhevsky A. (2014), One weird trick for parallelizing convolutional neural networks. URL: https://arxiv.org/abs/1404.5997.

[77] Brownlee J. (2017) Machine Learning Algorithms from Scratch: With Python. Jason Brownlee. URL: https://books.google.fi/books?id=ZTq2tAEACAAJ.

[78] Garcia-Garcia A., Orts-Escolano S., Oprea S., Villena-Martinez V. & Garcia-Rodriguez J. (2017), A review on deep learning techniques applied to semantic segmentation. DOI: https://doi.org/10.48550/arXiv.1704.06857.

[79] Sharma N. & Aggarwal L.M. (2010), Automated medical image segmentation techniques. DOI: 10.4103/0971-6203.58777.

[80] Ronneberger O., Fischer P. & Brox T. (2015), U-net: Convolutional networks for biomedical image segmentation. DOI: 10.48550/ARXIV.1505.04597.

[81] Hesamian M.H., Jia W., He X. & Kennedy P. (2019), Deep learning techniques for medical image segmentation: Achievements and challenges. DOI: https://doi.org/10.1007/s10278-019-00227-x.

[82] Rizwan I Haque I. & Neubert J. (2020) Deep learning approaches to biomedical image segmentation. Informatics in Medicine Unlocked 18, p. 100297. URL: http://www.sciencedirect.com/science/article/pii/S235291481930214X.

[83] Lai M. (2015) Deep learning for medical image segmentation. CoRR abs/1505.02000. URL: http://arxiv.org/abs/1505.02000.

[84] Hosny A., Parmar C., Quackenbush J., Schwartz L.H. & Aerts H.J.W.L. (2018) Artificial intelligence in radiology. Nature Reviews Cancer 18, p. 500 – 510. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85047143820&doi=10.1038%2fs41568-018-0016-5&partnerID=40&md5=ec52af34f46d0eac235babb471b093dd.

[85] Rantakallio P. (1988) The longitudinal study of the northern finland birth cohort of 1966. Paediatric and Perinatal Epidemiology 2, pp. 59–88. DOI: https://doi.org/10.1111/j.1365-3016.1988.tb00180.x.

[86] Burrus S.C. & Parks T.W. (1991) DFT/FFT and Convolution Algorithms: Theory and Implementation. John Wiley & Sons, Inc., USA, 1st ed.

[87] CVAT.ai Corporation (2022), Computer Vision Annotation Tool (CVAT). URL: https://github.com/opencv/cvat, Accessed: 24.1.2022.

[88] Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D. & Batra D. (2019) Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision 128, p. 336–359. URL: http://dx.doi.org/10.1007/s11263-019-01228-7, DOI: 10.1007/s11263-019-01228-7.

[89] Liszka-Hackzell J. & Martin D. (2002) Categorization and analysis of pain and activity in patients with low back pain using a neural network technique. Journal of medical systems 26, pp. 337–347. DOI: 10.1023/A:1015820804859.