Early Detection of Knee Osteoarthritis using Deep Learning-based MRI Features

Anastasis Alexopoulos

Delft University of Technology

Erasmus MC University Medical Center

TU Delft

Erasmus MC
University Medical Center Rotterdam

Early Detection of Knee Osteoarthritis using Deep Learning-based MRI Features

by

Anastasis Alexopoulos

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday August 23, 2022 at 10:00.

Student number:	4751345	
Project duration:	August 1, 2021 – August 23, 2022	
TU Delft Supervisor:	Prof. dr. ir. N. Tümer,	Delft University of Technology, Biomaterials & Tissue Biomechanics
EMC Supervisor:	Dr. J. Hirvasniemi,	Erasmus MC, Department of Radiology & Nuclear Medicine
Thesis committee:	Prof. dr. ir. N. Tümer,	Delft University of Technology, Biomaterials & Tissue Biomechanics
	Dr. J. Hirvasniemi,	Erasmus MC, Biomedical Imaging Group Rotterdam
	Dr. Q. Tao,	Delft University of Technology, Leiden University Medical Center

Abstract

***Background**: Advancements in the field of artificial intelligence have lead to the incorporation of automated algorithms in the analysis of medical images and data. Deep learning algorithms have been applied in musculoskeletal research to improve the understanding of osteoarthritis and to assist in disease detection and prognosis. The majority of the developed methods examine and process X-ray images and clinical data (age, gender etc.), with a small minority using MRI as inputs.*

***Objective**: The current master thesis project aims to investigate the influence of MRI scans on the early detection of knee osteoarthritis through the use of deep learning architectures, and to develop a semi-automatic method for knee region of interest extraction for creating the MRI input of detection algorithms.*

***Methods**: The MRI scans used in this project were acquired from the publicly available database of the Osteoarthritis Initiative. In total 593 dual echo steady state and intermediate-weighted turbo spin-echo sequences were included. The extraction of the knee joint included several processing steps. Initially, a U-Net model was trained on 507 annotated dual echo steady state MRIs for the segmentation of bone and cartilage tissue, which was followed by the registration of the output masks to intermediate-weighted turbo spin-echo sequences in order to create the joint labels for the desired MRI scans. Final step for the region of interest construction included the search of bone coordinates and the creation of the knee joint region of interest. The detection of early osteoarthritis progression from knee MRI scans was tested through three different deep learning architectures, a residual network (ResNet), a densely connected convolutional network (DenseNet) and a convolutional variational autoencoder (CVAE). Furthermore, the probability output of the ResNet and DenseNet as well as the feature vector of the CVAE were coupled with clinical data (age, gender, bone mass index) and used as input to a Logistic Regression Classifier, in order to investigate the influence of osteoarthritis related features to the detection task. The U-Net segmentation method was evaluated using Dice similarity coefficient and Intersection of Union while the detection algorithms using the area under the receiver's characteristic curve (AUC) and the precision-recall curve (PR-AUC) metrics, with two different input data configurations, only MRI and a combination of MRI and clinical data.*

***Results**: The U-Net algorithm for bone and cartilage segmentation showed adequate results, since Dice similarity coefficient and Intersection of Union reached mean values higher than 0.99 and 0.88. Regarding the early detection of knee osteoarthritis incidence, ResNet and DenseNet showed similar results, with both methods having an AUC value ranging from 0.5033 to 0.6269, when only MRI scans were examined. In the case of MRI and clinical data combination, the more complicated deep learning architecture (DenseNet) achieved the highest AUC at 0.6556. The best performing model was CVAE with the largest number of latent space features (1000) achieving an AUC of 0.6699 when combined with clinical data and an AUC of 0.6689 when used alone as input to the logistic regression classifier. All three deep learning algorithms yielded higher performance metrics when clinical data where combined with models' outputs.*

***Conclusion**: The tested deep learning algorithms showed a potential in the challenging task of early detection of knee osteoarthritis through MRI scans, even though they did not reach the same level of performance metrics. The region-of-interest creation had promising results for the implementation of U-Net method for bone tissue labelling.*

***Keywords**: Deep Learning, Knee, Osteoarthritis, MRI, Segmentation, ResNet, DenseNet, Autoencoder*

Acknowledgements

First and foremost, I would like to thank my main supervisor dr.ir. Nazli Tümer for her scientific guidance through my master years, her support and belief in me during this graduation project, no matter the difficulties and obstacles I faced until its completion.

I would also like to thank dr. Jukka Hirvasniemi, post doctoral researcher at Biomedical Imaging Group Rotterdam in Erasmus Medical Center and my daily supervisor, for sharing his experiences and knowledge in the field of deep learning and osteoarthritis and for helping me guide through the path of my master thesis project. His mentoring was a significant variable in my goal completion and I've learned many great things regarding scientific research and artificial intelligence from him during the past year.

Sincere thanks to my friends during my master studies, with whom we had amazing times.

Last but not least, deep-hearted thanks to my mother, Tereza, sister, Zoe, and my partner Elsa, for their love, support and belief in me.

Anastasis Alexopoulos
Delft, August 2022

Contents

Acronyms

AI artificial intelligence. 2

AUC area under the curve. 3

BICL Boston Imaging Core Lab. 7

BMI bone mass index. 4

BML bone marrow lesion. 2

DESS dual echo steady state. 2

DL deep learning. 1

FPR false positive rate. 3

FSE fast spin-echo. 5

IoU intersection of union. 3, 9, 29

IW intermediate weight. 7

KL Kellgren-Lawrence. 2

ML machine learning. 2

MOAKS MRI osteoarthritis knee score. 2

MRI magnetic resonance imaging. 1

MSE mean squared error. 14

NIfTI Neuroimaging Informatics Technology Initiative. 9

OA osteoarthritis. 1

OAI osteoarthritis Initiative. 4

PR precision recall. 4

ROC receiver operating characteristics. 3

ROI region of interest. 5

SE spin-echo. 2

SQ semi-quantitative. 7

TPR true positive rate. 3

TSE turbo spin-echo. 2

WORMS whole-organ MRI score. 2

1 Introduction

In the first sections of this chapter the main information regarding osteoarthritis (OA) and magnetic resonance imaging (MRI) modality role on OA diagnosis are presented, along with the basic principles of deep learning (DL) and its applications on OA detection and prediction. In the last sections the research aim, methodology and the outline of the report are stated.

1.1. Osteoarthritis

OA is a chronic joint disease that affects a significant portion of the world population. Although it can affect any synovial joint, the knee, hip and hand joints are the most common sites of OA development [1, 2]. OA is viewed as an age-related disease since it impacts a large portion of adults over 60 years of age and can lead to pain, stiffness and loss of mobility [1, 3]. OA is characterized by degeneration of articular cartilage and bone where the intrinsic repair mechanisms are insufficient. Joint space narrowing, osteophytosis, subchondral sclerosis, cyst formation and abnormalities of the bone contour are some of the radiographic features incorporated in the definition of OA[1]. Several elements have been identified to influence the development of OA in joints, that can be grouped into systemic and local risk factors. The systemic factors include age, gender, genetics, bone mass density, osteoporosis and nutrition, and may increase the joint susceptibility to injury or decrease the repair process in damaged bones and tissues. The local factors are thought to expose individual joints to injury and excessive loading, and may include obesity, joint deformities, muscle strength and weakness and acute injuries[1].

OA affects several structures of the joint, leading to articular cartilage loss, bone remodelling, synovitis and lesion development in the bone marrow[4]. These outcomes of OA pathogenesis influence the joint in a nonuniform way since areas with cartilage loss and bone deformities can further increase mechanical stresses and cause joint misalignment, which lead to higher loading that deteriorates joint integrity, creating a cycle of joint degradation. Furthermore, areas of the synovium and the cartilage with inflammation can contribute to joint pain and degeneration.

The structural changes in the joint compartments are the basis of OA detection and diagnosis, along with the presence of symptoms such as joint stiffness and pain, even though they are detected when the disease is advanced and irreversible[3, 4]. The early identification of cartilage loss or subchondral bone degeneration on patients without or few symptoms can assist the treatment effectiveness and the disease progression[3]. Several imaging modalities are used to depict the joint structures. The traditional method used is X-ray images which are created in a few seconds and allow the extraction of 2D morphological and statistical features, however hinder the detection of early developed OA measurements, such as localised cartilage degradation, due to them being 2D projections of 3D structures[3, 5]. A more modern technique is MRI modality that reveals high resolution

3D structures and has higher sensitivity in detection of early joint changes. Furthermore, MRI does not involve ionising radiation but may cause claustrophobia to some patients and excludes people with implants [3, 5].

The most commonly used method to assess the severity of OA is the Kellgren-Lawrence (KL) grade scale[6] which is based on X-rays. The Kellgren-Lawrence classification system is the first method for assessing the severity of joint OA. It consists of five grades based on the presence of OA features, from 0 for no joint space narrowing or reactive damage to 4 for large ostephytes, severe joint space narrowing and sclerosis and definite bone ends deformity[6].

Two semi-quantitative grading systems, that are widely used, are based on MRI scans, whole-organ MRI score (WORMS)[7] and the MRI osteoarthritis knee score (MOAKS)[8]. WORMS evaluation incorporates T1- and T2-weighted spin-echo (SE) sagittal and coronal MRI scans with and without fat-suppression sequences, while MOAKS mostly uses 3D T1-, T2-, intermediate-, proton-density-weighted fat suppressed turbo spin-echo (TSE) and dual echo steady state (DESS) sequences. In both scoring methods, features such as bone marrow lesion (BML) size, percentage of cartilage thickness loss, osteophytes formation size, synovial volume, meniscal and ligament abnormalities are graded, using different scales for each feature. The main difference between WORMS and MOAKS is that the latter has implemented a detailed assessment of BML regions and scores and omitted redundancy information regarding cartilage[8].

1.2. Magnetic Resonance Imaging

MRI modality is considered by many one of the most informative imaging modalities regarding the structure of the knee joint [9, 10]. The key aspect lies in the ability of MRI scans to provide detail depictions of soft tissues, allowing all joint compartments to be examined simultaneously. Furthermore, the direct construction of 3D images offers a multiplanar tomographic view, which excludes projectional distortion, magnification and overlapping structural superimposition occurring in 2D X-rays. Another advantage of using MRI scans in OA detection is the examination both the morphology and the compositional parameters closely linked to arthritic processes. Thus, its ability to create relatively high spatial resolution visualizations of subchondral bone, cartilage, ligaments, synovium and meniscus makes the MRI modality more suitable for the early detection of knee OA when compared to X-rays.

1.3. Deep Learning

A scientific field that is being applied more frequently in the musculoskeletal imaging and other applications, due to the growth in computing power, is artificial intelligence (AI), which is involved with aspects and terms from pattern recognition, probability theory and statistical analysis. AI is composed of two main groups, machine learning (ML) and its subcategory, DL, and aims to automatically extract patterns in data. Their main difference is that deep learning methods learn patterns directly from their input images, while in machine learning the input data should be predefined by the developer of the method. So it can be said that machine learning is feature-based while deep learning is input-based[11].

Deep learning methods can perform two types of tasks, classification and regression. Classification categorizes the input data into a particular class according to the presence of a specific feature, for example the classification of osteophyte formation in OA, and its output is a discrete value. On the other hand, regression problems map the input to a continuous value, such as the staging of OA. A subcategory of classification task is image segmentation DL algorithms, which aims to divide regions of interest within an image by classifying each pixel as being part of a specific tissue compartment. In order to perform the classification and regression tasks, two main strategies are applied in the algorithms. Supervised learning which requires the combination of data with labels that act as the ground truth, while unsupervised learning does not require the use of labelled data. The process through which the performance of a deep learning model is optimized is called model training and, in most cases, aims to minimize the error between the model's output and the ground truth, also known as loss function[12–15].

Convolutional Neural Networks

Deep learning algorithms are based on neural networks, which are composed of several connected nodes, called neurons, that construct a layer. A node receives the input which is combined with weights, sums it up and with the application of an activation function creates the output. One of the most notable classes of deep learning models in image processing tasks is convolutional neural networks [13] with the following components:

Convolutional Layers

Convolutional layers utilize various kernels that slide over the whole image (convolve) and intermediate feature maps, save the operation result at each position and generate the output feature map of the layer. The kernels are weight matrices designed to detect specific patterns in the input image. The advantages of convolutional layers include the reduced number of parameters due to weight sharing, the invariance to object position and the correlations of close pixels due to local connectivity of the kernels[13]. The convolutional function can be depicted with the following expression:

$$\mathcal{O}(i,j) \equiv (K * I)(i,j) = \sum_{m}^{M}\sum_{n}^{N} I(i+m, j+n)K(m,n)$$

With I being the input, K the kernel, M, N the size of the kernel and $\mathcal{O}$ the output feature maps.

Pooling Layers

Pooling layers usually are placed after convolutional layers and reduce the dimensions of the feature maps and the number of network parameters. Similar to convolutional layers, pooling layers are translation invariant since the pooling function slides over a specific number of pixels until it covers the whole image. The two most commonly used types of pooling functions are averaging and maximizing, with the first keeping the average and the second the maximum value of neighboring pixels.

Fully-Connected Layers

Fully connected layers are usually used in the last part of the convolutional network, their containing nodes are linked with each node of the previous layer and their output further incorporated for the network's task. The disadvantage of these layers is their large number of parameters which leads to higher requirements in computational power for their training.

Performance Metrics

The performance of deep learning models for the detection, prediction and severity classification of knee OA can be measured with the use of the confusion matrix (table 1.1) from which several performance metrics (table 1.2) can be extracted, such as accuracy, sensitivity, specificity and precision. The segmentation of a medical image is measured through the intersection of union (IoU) and Dice Similarity Coefficient metrics (table 1.2). The columns of a confusion matrix represent the prediction results and the rows the ground truth labels.

		Prediction	
		1 (with OA)	0 (without OA)
Ground Truth	1 (with OA)	True Positive (TP)	False Negative (FN)
	0 (without OA)	False Positive (FP)	True Negative (TN)

Table 1.1: Confusion Matrix

The area under the curve (AUC) receiver operating characteristics (ROC) can be used to evaluate the ranking performance of a classifier. The axis of the graph are usually two different pairs of metrics, Sensitivity & (1-Specificity) and true positive rate (TPR) & false positive rate (FPR) at the Y and X axis. The points (0,0) and (1,1) of a ROC curve depict the training-free classifiers *Always Negative* and *Always Positive* respectively, while

Metric	Formula	Evaluation Focus
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$	Ratio of correct predictions over the total number of instances evaluated
Sensitivity or True Positive Rate	$\frac{TP}{TP+FN}$	Fraction of positive instances that are correctly classified
Specificity or True Negative Rate	$\frac{TN}{TN+FP}$	Fraction of negative instances that are correctly classified
Precision	$\frac{TP}{TP+FP}$	Ratio of correctly predicted positive instances from the total predicted positive instances
Intersection of Union	$\frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$	Measures the similarity and diversity between sample sets
Dice Similarity Coefficient	$\frac{2 \times \lvert A \cap B \rvert}{\lvert A \rvert + \lvert B \rvert}$	Measures spatial overlap between two segmentations

Table 1.2: Evaluation metrics

the point (1,0) shows the ideal classifier and (0,1) the classifier that outputs always wrong results. The ROC graph is divided diagonally by the (0,0)-(1,1) line which represents a non-discriminative algorithm (*TPR=1-FPR=TNR*). In the upper left part of this graph are the classifiers that perform better than random and on the lower right part those who perform worse than random.

Apart from AUC ROC performance analysis, another common metric for binary classification models is the precision recall (PR) AUC which shows the precision (positive predictive value) as a function of recall (true positive rate). In PR-AUC graphs, precision is placed in Y axis while recall in X, and the curve is maximized in the upper right corner.

1.4. Deep Learning Methods In Knee Osteoarthritis

The majority of the developed deep learning algorithms for the detection, severity diagnosis and progression of knee OA use X-rays as their input, mostly from the publicly available database of osteoarthritis Initiative (OAI). In two research papers conducted by Tiulpin[16, 17], the ability of a residual based network to grade OA from knee X-rays was tested, with the first one[16] implementing a Siamese architecture that reached an AUC of 0.93 for the detection of subjects with KL$\geq$2, and the second one[17] examining an ensemble of residual networks for the automatic grading of knee OA that achieved an average accuracy of 66.68% and an AUC of 0.98 for OA vs no-OA classification. Another publication of Tiulpin[18], utilized X-rays as the input to a residual based network, combined its output predictions with clinical data (age, gender, bone mass index (BMI), symptomatic assessment) and reached an AUC of 0.81 for the task OA progression prognosis. The influence of clinical data on the detection of OA severity was also studied by Kim[19], which showed that combination of the output predictions of a residual-based network with clinical data can improve the classifier's AUC value for KL grades 0, 1 and 2.

In a research paper published by Norman[20] the classification of OA grade severity was implemented through the use of an ensemble of different versions of DenseNet and demographics and reached sensitivity values of 0.86, 0.69 and 0.70 for KL grades 4,3 and 2 respectively. A version of DenseNet with a larger number of layers for the OA severity detection task was developed by Thomas[21], where the network that was trained with both original and augmented X-ray images yielded an average accuracy of 0.71 for each KL grade and 0.87 for OA vs no-OA KL grades. The ability of a DenseNet to predict OA progression from X-rays was tested by Guan[22]. The best performing developed method combined the predictions of the DenseNet with clinical and radiographic risk factors and yielded an AUC of 0.86.

A different deep learning algorithm was examined in the paper published by Nasser[23] for the detection of knee OA. An autoencoder was developed that received as input knee X-rays, extracted several discriminative features and used different classifiers to distinguish OA vs no-OA subjects, reaching an accuracy of 0.83. In a paper published by Chen[24] X-ray images were tested by several deep learning architectures for KL grade classification with the VGG-19 method reaching the highest average accuracy at 0.70.

Two reports that examined MR imaging modalities as inputs for the detection of knee OA were published by Pedoia[25, 26]. The performance of a shallow 3D convolutional network for the detection of lesions in the meniscus and patella from 3D fast spin-echo (FSE) MRI sequences was tested[25], reaching an AUC of 0.89 and 0.88 for meniscus and patella lesions respectively. In the second publication,[26], T2-weighted MRI sequences were used as input to a DenseNet method, with the combinations of its predictions with clinical data achieving an AUC of 0.82 for detecting knee OA.

Another study published by Tolpadi [27] investigated the ability to predict total knee replacement from MRI scans within 5 years. In this publication, the tested DenseNet incorporated X-rays and 3D IW-TSE MRI sequences along with clinical data, and compared their performance metrics. The results revealed that in MRI based model performed similarly to the X-ray based model on predicting total knee replacement with AUC 0.89 and 0.83 respectively, when imaging and clinical data were combined, and MRI based model outperformed X-ray based model when used as input clinical and imaging data from no-OA subjects at baseline reaching an AUC of 0.943 vs 0.799. In the research conducted by Schiratti [28] the performance to predict the progression of OA based on joint space narrowing within 12 month from MRI scans was investigated. 2D IW-TSE and 3D DESS MRI sequences along with clinical data on subjects' experienced pain were used as input and the tested EfficientNet achieved an AUC of 0.65 and 0.63 when 2D IW-TSE and 3D DESS MRI scans respectively.

1.5. Thesis Objective

A large number of scientific publications regarding knee OA were mainly focused on the disease severity classification from X-ray images, with a small portion of the existing literature examining the ability of deep learning algorithms for the early progression prognosis. MRI modality has not been yet investigated thoroughly for it's impact on the prediction of knee OA development, even though this type of imaging modality can detect bone and cartilage changes that may increase the risk of joint collapse at short periods of time (1-2 years)[29, 30]. Furthermore, the majority of the deep learning methods developed for the detection of knee OA are based on ResNet and its more elaborate version, DenseNet, with a few reports investigating the performance of other methods for OA diagnosis and prognosis task.

The aim of the current master thesis project can be encapsulated by the following research question:

Do deep learning-based MRI features influence the early detection of knee osteoarthritis progression?

The examination of this research question raised one significant challenge which is listed below as a subsidiary questions:

- How accurate is the automatic extraction of regions of interest from knee MRI scans?

1.6. Research Methodology

The aim of the current master thesis project is to examine the ability of deep learning algorithms to detect the early progression of knee OA between baseline and 24 months follow-up period through MRI scans. For this purpose a set of medical knee MRI images were collected from the publicly available database of the OAI[31]. With respect to knee OA progression detection, the images were categorized into two different groups: (i) knee with OA and (ii) control (knee without OA) based on presence of OA features on knee MRIs after 24 months.

The second step in the methodology of this work was to segment the MRI scans into the knee joint region. For this process, a 2D segmentation method (U-Net[32]) was applied in order to identify the different bone regions and thus extract their pixel coordinates. This allowed for the extraction of the minimum and maximum coordinates of the tibial and femoral bones which were used to create the region of interest (ROI) around the knee joint.

Having identified the size of the cropping area, the third step was to create the 3D knee MRI scans for both the control and progress groups and use them as input to investigate deep learning algorithms for the early diagnosis of OA progression. A review of the existing work and literature related to OA prognosis and classification reveals a large number of techniques with no clear ”best” model suggestions. Thus, in the current thesis project three approaches were examined, two that most commonly applied in previous publications, a Residual Network (ResNet[33]) and a DenseNet[34], and a not so commonly used method, a Convolutional Autoencoder[35].

1.7. Thesis Organisation

The rest of the master thesis is arranged as follow. Chapter 2 provides more information regarding the materials and methods applied in this project, the process the dataset was created, the ROI extraction procedures of the MRI scans, that involve bone tissue segmentation and different MRI sequences registration, and the deep learning architectures tested for early detection of knee OA. The results of the developed semi-automated ROI extraction and detection algorithms are presented in Chapter 3, discussion of the main findings in terms of the research question in Chapter 4 and the thesis conclusion in Chapter 5.

2

Materials and Methods

The procedures and the developed deep learning methods used along the project are described in detail in the current chapter. These include the dataset creation, the choice of MRI sequences as well as the segmentation of the knee joint area. Furthermore, an elaborate presentation of the different deep learning methods that were tested for the early detection of knee OA progression is included.

2.1. MRI Data Description

The data used for this research were retrieved from the publicly available database of OAI. The OAI[31] provides data from a multi-center, longitudinal, prospective observational study of knee OA. The OAI cohort includes 4796 participants of both genders between 45 and 79 years of age and consists of medical images, clinical data and biospecimens, collected at baseline and at four follow-up visits (12-,24-36-,48- months). MRI images of the OAI datasets were acquired using Siements 3 Tesla scanners, with a total acquisition time of 75 min for both knees. The MRI sequences examined in this master thesis project are the coronal intermediate weight (IW) 2D turbo spin-echo (TSE) (COR 2D IW-TSE) and the sagittal 3D dual-echo in steady state (DESS) with water excitation (SAG 3D DESS WE), with total acquisition time 6.8 and 21.2 minutes respectively. DESS sequence is used for the segmentation of the knee joint due to the utilization of already annotated DESS scans and IW-TSE is used for the early detection of knee OA, due to it's sensitivity and specificity to bone and cartilage changes. The MRI protocol acquisition parameters for the two examined sequences are shown in table 2.1:

Scans	Plane	FS	Matrix (phase)	Matrix (freq)	No. of slices	FOV (mm)	Slice thick./gap (mm/mm)	Flip angle (degrees)
COR IW 2D TSE	Coronal	No	307	384	35	140	3/0	180
SAG 3D DESS WE	Sagittal	WE	307	384	160	140	0.7/0	25

Table 2.1: Acquisition parameters for coronal 2D IW-TSE and sagittal 3D DESS.
FS: fat suppression, WE: water excitation, FOV: field of view

The two OAI datasets were chosen to be examined for the task of early detection of OA through knee MRIs are the knee MRI (kMRI) SQ MOAKS (Boston Imaging Core Lab (BICL)) and the knee MRI (kMRI) SQ WORMS datasets, that reported centrally performed longitudinal semi-quantitative (SQ) readings of OA from knee MRI performed at the BICL. The first dataset contains data from five projects with knee joint changes assessed by MOAKS [36] grading scale, however we examined projects 22 and 65 with 600 and 1033 subjects at baseline respectively due to the aspect that they were the only projects with data from five time points (baseline, 12-, 24-, 36-, 48-month visits). The second dataset used the same MRI sequences as the first one, from three time points (baseline, 24-months, 48-months) and the structural abnormalities of the knee joint were assessed by using the WORMS[7] system.

The two above mentioned grading systems assess the presence of several OA related features such as degenerative changes in cartilage, bone marrow lesions, meniscus and formation of osteophytes. According to WORMS, OA in the whole knee joint is diagnosed if the following items are fulfilled [7]:

- Cartilage morphology score $>= 3$
- Bone marrow lesions (BML) score $>= 2$
- Osteophytes $>= 2$

The MRI definition of tibiofemoral OA according to MOAKS grading scale as defined by Hunter [37] is based on the presence of both group [A] features or one group [A] and two or more group [B] features.
Group [A]:

- Definite osteophyte formation $>= 2$
- Full thickness cartilage loss $>= 3$

Group [B]:

- Subchondral bone marrow lesion or cyst $>= 1$
- Meniscal subluxation, maceration or degenerative tear $>= 2$
- Partial Thickness cartilage loss (where full thickness loss is not present) $3 >=$grade$>= 1$

The first step was to apply a feature based search in the datasets in order to find which subjects had the above features reported and at which follow-up visits. Furthermore, the focus was on which subjects had features that exceeded the OA presence threshold as mentioned above, thus showing the progression of OA between specific visits. This initial investigation of the datasets revealed several issues regarding the formulation of MRI OA definition. The main aspect was that both datasets (kMRI SQ MOAKS (BICL), kMRI SQ WORMS) did not report the osteophytes score in the majority of the subjects and not in every time point, even though the scores of BML, cartilage thickness loss and meniscus degradation were. Another question that was raised was whether it is possible to combine the two datasets even though the same grading scale was not applied to assess the state of OA features in MRI scans, and which time points should be chosen in order to define the early OA progression.

For the lack of osteophytes scores, it was decided based on the existing literature to extract the participants' osteophyte grade from their X-rays. Several research publications[38–40] highlight the correlation between osteophytes depicted in radiographs and cartilage damage in MRI, which supports the role of osteophytes in OA process. With the incorporation of X-ray detected osteophytes, the OA presence threshold needed clarification since KL grade 1 was considered as a possible osteophytic lipping[6], thus based on previous research papers[41] subjects with KL grade 1 should be treated as early OA group. Regarding the issue of using data from both studies, even though they apply different OA grading systems (WORMS, MOAKS), it was decided to include both datasets for the final control and case groups creation, since MOAKS is a newer and refined version of WORMS as stated in several publications[37, 42]. Both groups were composed of participants from two follow-up visits (12-, 24- months) due to the decreasing number of subjects without OA in later time points.

As a result of the above mentioned challenges regarding the two datasets, the distinction between subject with and without OA progression within the first two follow-up visits was accomplished by applying a modified MOAKS grading system that included thickness cartilage loss, BML and meniscal degradation grades from MOAKS and WORMS and osteophyte formation grades from KL method. Those OA variables we extracted from the OAI database documents of the kMRI SQ MOAKS (BICL) and kMRI SQ WORMS studies. The final number of subjects that are comprise control and case groups is shown in section Control and Case groups of chapter Results.

Based on the identification number of the subjects included in the final control and case groups, the clinical data along with the sagittal 3D DESS WE and coronal 2D IW-TSE MRI sequences from baseline time point were acquired from the OAI database. The DESS sequence was used in order to segment the bone tissue for the ROI extraction and the IW-TSE images were the input image of the OA detection algorithms, due to their sufficiency in finding central cartilage damage[8] and BMLs[43] when compared to DESS. Furthermore, clinical information (age, gender, BMI) were combined with the output of the deep learning methods to create the input dataset of the logistic regression classifier.

2.2. Extraction of region-of-interest

In order to create the area around the knee joint, a semi-automated extraction method was developed, that involved the identification of the pixels that belonged to the tibial and femoral bones and the acquisition of the minimum and maximum coordinates of those pixels. This processing pipeline incorporated the DESS sequences for bone mask creation, the IW-TSE images registration to the DESS masks and the bone coordinates labelling in the desired MRI sequence.

2.2.1. DESS Segmentation

The segmentation of DESS images was achieved with the application of U-Net method[44] which is constructed of four down-sampling and up-sampling steps in order to extract context and spatial information, where each pixel is and in which label it belong to (tibia, femur). Each down-sampling step was composed of two 2D convolution layers, one dropout and one max-pooling layer. Each up-sampling step included one transpose convolution, one dropout and two convolution layers, as 2.1 shows. The tested segmentation method was based on a previously built U-Net model during a master student's internship project in collaboration with Erasmus Medical Center and Delft University of Technology and was trained with the use of Adam's optimizer, categorical cross-entropy loss function and its performance was assessed using the Dice Similarity Coefficient, mean IoU and pixel classification accuracy metrics. Dice Similarity Coefficient shows the spatial overlap between two segmentations while IoU measures the accuracy of segmentation by calculating the fraction of the area of overlap between ground truth and predicted regions over the area of union between the same regions.

For the training of the segmentation model, the publicly available by Ambellan[45] dataset was used, which contains 507 3D DESS MRI scans with the same acquisition parameters as mentioned in table 2.1. The slice annotation performed by the experts concerned the tibial and femoral bone and cartilage tissues of the sagittal plane, thus the U-Net was trained and validated with 81,120 slices. The results with the specific hyper-parameters' values are shown in chapter Results, section U-Net. After the training and validation process, the developed 2D U-Net received as input the 593 DESS MRI scans of both the control and case groups in order to create the labelled masks, that classify each pixel to five classes, 0 for background, 1 and 3 for femoral and tibial bones respectively, and classes 2 and 4 for femoral and tibial cartilage tissue respectively. This resulted in 95,040 labelled slices of both control and case groups.

2.2.2. DESS to IW-TSE Registration

Since the MRI sequence that was used for the detection of early OA progression was IW-TSE, the registration of DESS to IW-TSE sequences was needed in order to create the IW-TSE masks and extract the knee joint ROI. The pipeline for the acquisition of the IW-TSE labelled images was applied in the 3D Neuroimaging Informatics Technology Initiative (NIfTI) files of both sequences, incorporated with the use of ElastiX[46], a medical image registration library, is presented in figure 2.2 and can be described with the following steps:

1. The DESS MRI scans for each patient with 384×384×160 slices (axial, coronal, sagittal planes) were registered through an affine transformation into IW-TSE sequences with 384×37×384 slices (axial, coronal, sagittal).

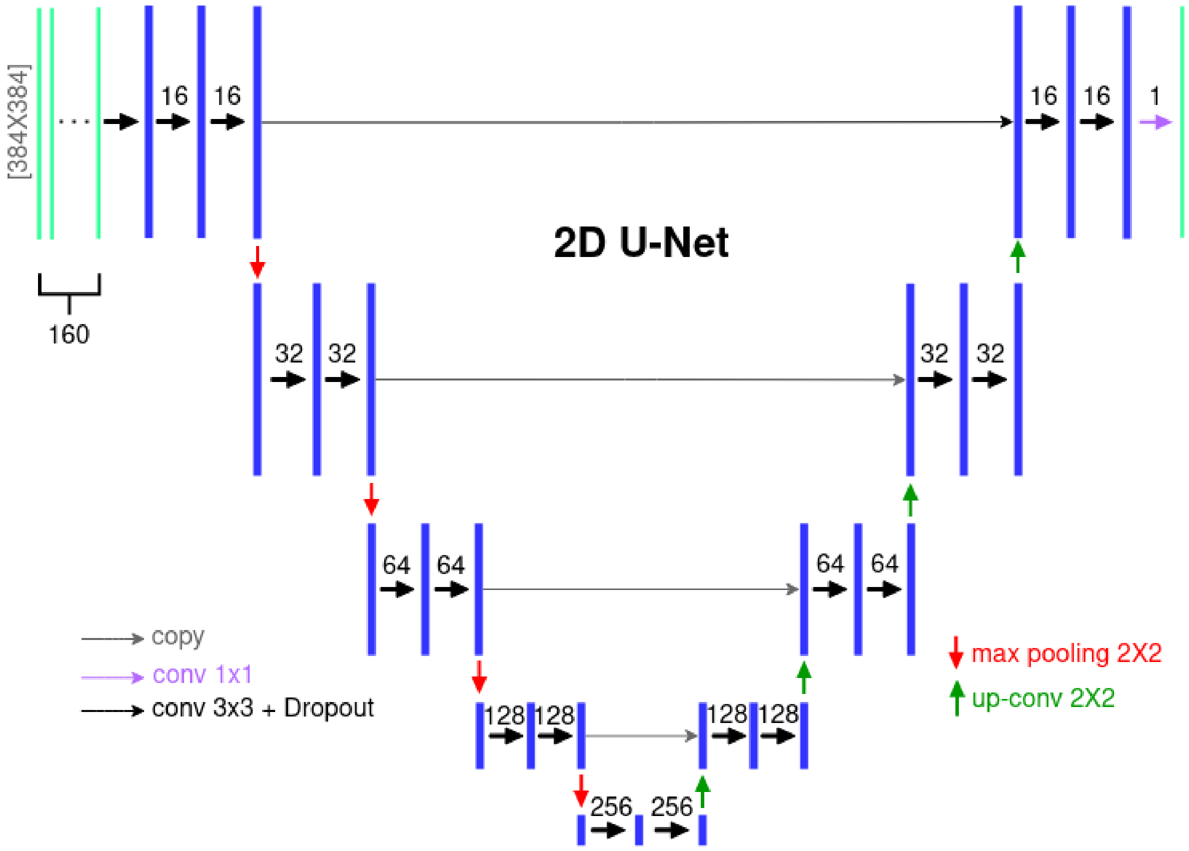


Figure 2.1: 2D U-Net for segmentation of sagittal 2D DESS MRI scans for bone and cartilage mask creation[45]

2. The initial image registration gave two outputs, a new image that combined the two MRI sequences and the transformation parameter map that created for this registration.

3. With the use of the transformation parameter map, the DESS masks created by the U-Net were registered to the combined DESS/IW-TSE image and the desired IW-TSE bone segmentation of 384×37×384 slices (axial, coronal, sagittal planes) was created.

The quality of registration was evaluated qualitatively through a visual inspection of the superimposed target IW-TSE image and the registered IW-TSE mask. Out of the 593 registrations only one failed to create a proper labelled IW-TSE image, and after the application of different registration methods with the same outcome, it was discarded.

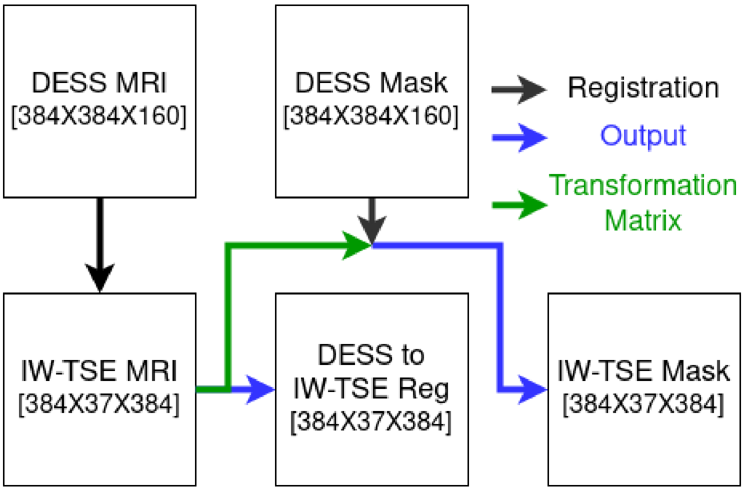


Figure 2.2: Image Registration Pipeline
for the creation of IW-TSE bone and cartilage masks

2.2.3. Bone Tissue Coordinates

The next step for the creation of the knee ROI was the search of minimum and maximum pixel coordinates of the tibial and femoral bones. Due to the imperfections of the segmentation and registration processes, several pixels were misclassified and instead of belonging into the background class indicated bone areas. For the purpose of discarding the misclassified areas, the Canny Edge detection algorithm[47] was applied in each slice in the created IW-TSE masks. The distinction between those bones during the searching process was accomplished through the use of their label values from the IW-TSE mask image (femur: 1, tibia: 3). The IW-TSE MRIs and their corresponding masks were flipped so each knee joint (left or right) had right knee orientation and after a global search in both datasets (control and case) the minimum and maximum coordinates were found. The implementation of the Canny Edge Detection pipeline is presented in figure 2.3

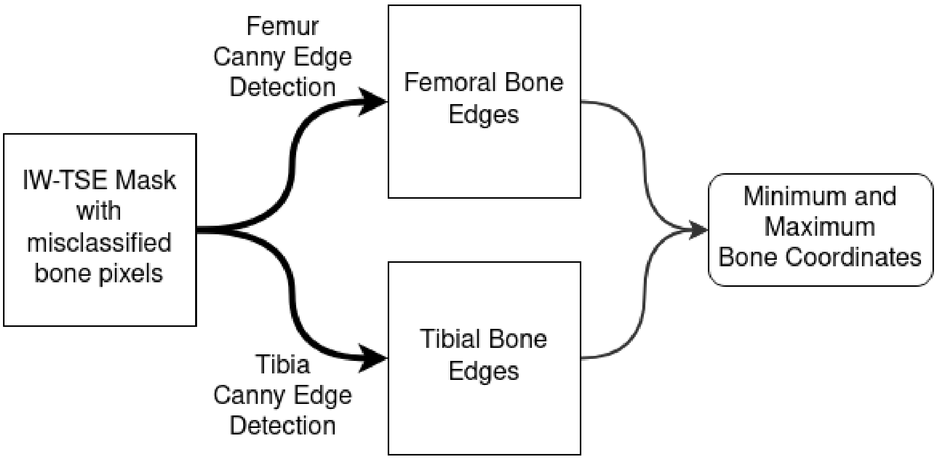


Figure 2.3: Canny Edge Detection Pipeline
Femur variance σ, low threshold$=0.37$, high threshold$=0.39$
Tibia variance σ, low threshold$=0.90$, high threshold$=0.95$

2.2.4. 3D IW-TSE input

The next step in the ROI extraction pipeline is the selection of those coronal slices that depict bone tissue and not only soft tissue or image background. For this purpose, a condition on the number of bone pixels present in a slice was implemented. After a visual inspection of the results with different threshold values, the final applied threshold was the following: # of bone pixels in each slice > 500 pixels.

By setting the above threshold, the result was the reduction of coronal plane slices in the initial 3D IW-TSE MRI scans. Instead of the initial 37 coronal slices, the new images varied from 18 to 27 coronal slices. In order to standardize the shape of the image inputs for the detection algorithms, the interpolation of those 3D images with coronal slices greater than 18 was applied through the use of *ndimage.zoom* function of SciPy Python image processing library.

After the application of slice interpolation, the knee joint regions were extracted by applying a cropping function with the desired rectangular shape. The final 591 3D IW-TSE MRI input scans that composed the control and case groups had the following shape: 250X18X320 slices in the axial, coronal and sagittal planes respectively.

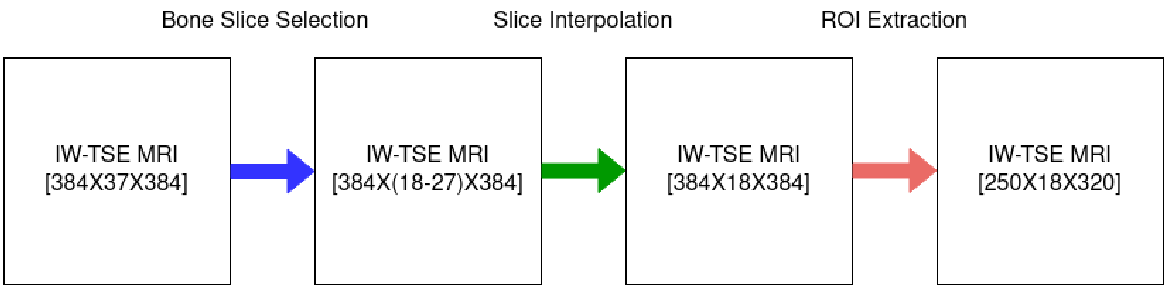


Figure 2.4: Flowchart for the creation of final 3D IW-TSE MRI input (axial, coronal, sagittal)

2.3. Detection Algorithms

Several deep learning architectures have been tested for their ability to classify the severity of OA or to detect the presence of OA in knee imaging modalities, both X-ray images and MRIs, by processing 2D, 3D inputs in combination with clinical information, as mentioned in section 1.4. The most commonly used according to the existing literature include ResNet[33] and DenseNet[34], which were tested in the current master thesis project. Along with these two deep learning methods, an additional algorithm was chosen to be examined for the task of early detection of knee OA through MRI scans, a Convolutional Variational Autoencoder[48], which be applied for binary classification and its performance with MRI scans as input can be compared to existing publication that used X-ray images as input[23].

In the following subsections the implementation details of the above mentioned deep learning algorithms for the classification task of early detection of knee OA through knee MRIs are presented.

2.3.1. Residual Network (ResNet)

Deep residual networks are based residual learning which instead of aiming that few stacked layers fit a desired mapping $\mathcal{H}(x)$, aim at a residual mapping $\mathcal{F}(x) := \mathcal{H}(x) - x$. Thus the original function becomes $\mathcal{F}(x) + x$, with $+x$ showing a shortcut connection performing an identity mapping. This combination of residual and identity mapping forms a residual block, the building element of residual networks, that can be expressed by the following function: $y = \mathcal{F}(x, W_i) + x$, with x and y being the input and output respectively. Several variations of ResNet have been developed, the shallower versions (ResNet-18, ResNet-34) with two 3×3 convolutional layers in the residual blocks and the deeper versions (ResNet-50, ResNet-101, ResNet-152) with two 1×1 and one 3×3 convolutional layers. The two residual blocks can be seen in figure 2.5.

Due to the limited computational resources, the ResNet-50 was chosen for this master thesis project, which is presented in figure 2.6. ResNet-50 is composed of 50 layers, of four different residual block, with each block repeated several times (3,4,6,3). Each repetition of the residual blocks is applied with different number of filters and its concluded with compressing the output feature map to match the input dimensions of the next repetition. The implemented ResNet-50 was modified in order to be able to receive the desired 3D IW-TSE MRI inputs. The total number of learned parameters was around 45 million and was trained with the use of binary cross-entropy

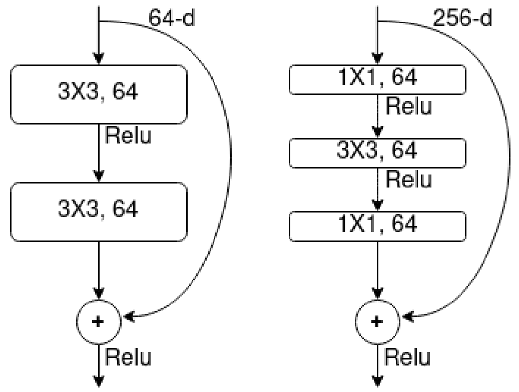


Figure 2.5: Residual Blocks. Left: for ResNet-18/34, Right: for ResNet-50/101/152

function, giving as an output the probability of a single 3D image belonging to the progress group.

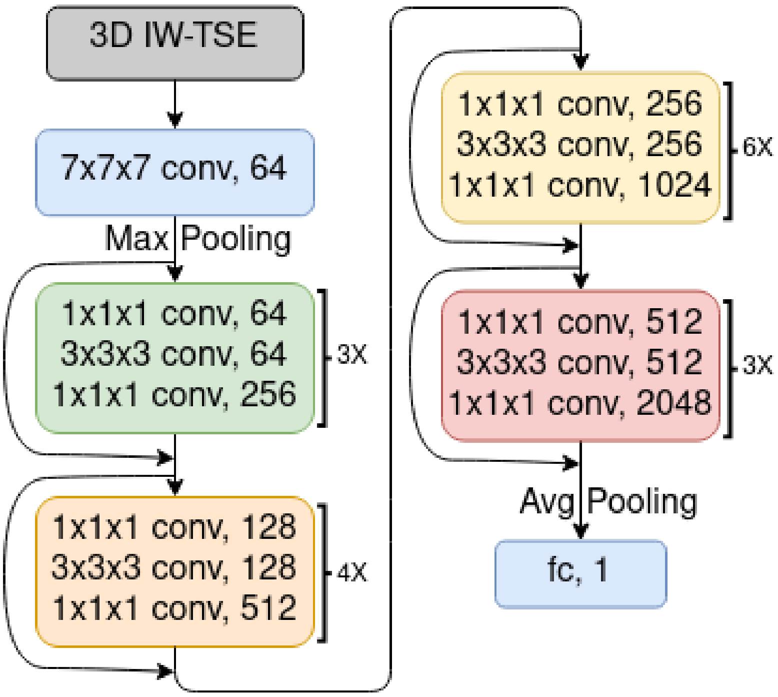


Figure 2.6: ResNet-50

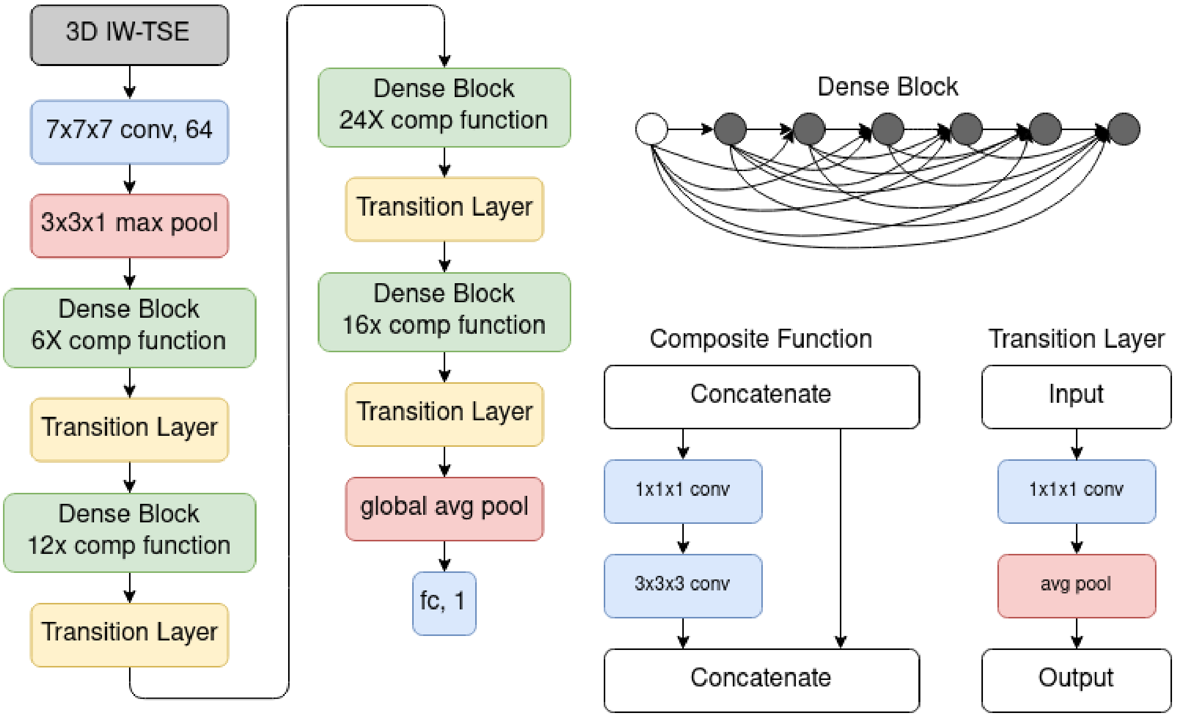
2.3.2. Densely Connected Convolutional Network (DenseNet)

DenseNet constitutes a more elaborate and complicated version of Residual Network. DenseNet utilizes the outputs of several convolutional layers through features reuse, thus resulting in an easier-to-train and parameter-efficient network. The main difference between DenseNet and ResNet is the concatenation of feature-maps learned by all previous layers, which increases the variation in the input values of subsequent layers.

The connectivity of DenseNet can be shown by the following input expression of the l^{th} layer: $x_l = H_l([x_0, x_1, ..., x_l])$, where $[x_0, x_1, ..., x_{l-1}]$ refers to the concatenation of features produced by the previous layers. The building operations of the DenseNet are called composite function and include of three layers, a batch normalization, followed by a rectified linear unit (ReLU) and a convolutional layer. Two composite functions, with 1×1 and 3×3 convolution kernel sizes respectively, construct the Dense Block. Four dense blocks are used to create the DenseNet architecture with specific filters and repetitions, which lead to DenseNets with different depth (121, 169, 201, 264 layers). Between each of these blocks a transition layer is implemented, which facilitates the down-sampling of feature maps and thus the networks parameter efficiency.

The shallower version of DenseNet, DenseNet-121, was tested in this project. This model version is composed of 121 layer, of four dense block with different feature maps and repetitions (6, 12, 24, 16) for each block. The parameters of the convolution and pooling layers were modified in order to be able to receive the 3D IW-TSE MRI inputs, resulting to a total number of around 11 million trainable parameters. DenseNet was also trained with the incorporation of binary cross-entropy loss function.

In the figure 2.7, a schematic depiction of the developed DenseNet-121 for the task of early detection of knee OA through MRI scans can be seen:



(a) DenseNet-121 **(b)** DenseNet-121 building blocks

Figure 2.7: 3D DenseNet-121.
a) A detailed depiction of the developed DenseNet-121
b) First dense block with 6 repetitions of composite function (upper)
Composite Function with two convolution layers,
Transition Layer with one convolution and one average pooling layer (lower)

2.3.3. Convolutional Variational Autoencoder (CVAE)

The autoencoder is an unsupervised learning algorithm that has as a goal to reconstruct the output from the input image. It consists of an encoding part, which is employed to encode the input image into a latent-space representation, and a decoding part, which reconstructs the encoded features. In order to encode the input images, the encoder maps this input data through the use of several convolutional or fully connected layers. The created final representation can be depicted with the following function $z = f_e(W_e^L \times h_{L-1} + b_e^L)$ where f_e is the activation function, h_{L-1} shows the output of the previous layer, L the current layer, W_e^L and b_e^L the weight and bias terms of the specific layer. To obtain the reconstructed image, the decoder acts in a similar way as the encoder, with the feature code being its input. The original form of the autoencoder uses the mean squared error (MSE) as a loss function for its optimization process.

A method different from the traditional autoencoder was tested for the task of early detection of knee OA, called Convolutional Variational Autoencoder (CVAE), based on existing publications [23, 35]. CVAE is constructed with the use of 3D convolutional and de-convolutional layers for the encoding and decoding part respectively, and the addition of latent space reparameterization. This code reparameterization enables the feature vector to be determined by a multivariable Gaussian distribution $z = \mu_x + \sigma \times \epsilon$, where $\epsilon \sim N(0, I)$, with μ being the mean, σ the standard deviation and ϵ Gaussian Noise variable. Therefore, in variational autoencoder, the encoder outputs a probability distribution in the bottleneck layer, from which the latent space is sampled.

Another change that was examined in the tested CVAE was a modified reconstruction loss function. Due to the complex nature of knee MRI scans and the high similarity between ROIs of healthy and early OA progression subjects, a discriminative term was incorporated in the object function, based on the work of Nasser et.al.[23], which forces the network to extract those features that minimize intra-class and maximize inter-class distances. This term is called discriminative penalty and is expressed by the following function: $\Omega_{disc} = \frac{\sigma_1^2+\sigma_2^2}{|\mu_1-\mu_2|^2}$, where μ_i and σ_i^2 are the mean and variance of the learned latent space respectively of each class. Thus, the overall loss function can be expressed with the following terms: $J_{CVAE} = J_{MSE} + \lambda\Omega_{disc}$, where λ is the discriminative penalty weight. The developed CVAE is constructed with three convolution and three de-convolution layers in the encoding and decoding part, along with two dense layers for the creation of the latent space, resulting in around 22 million parameters. In the figure 2.8 a schematic depiction of the models' architecture is shown.

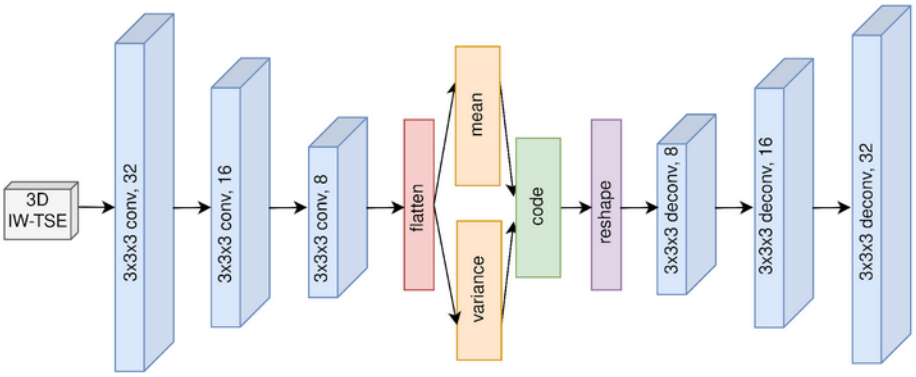


Figure 2.8: 3D Convolutional Variational Autoencoder

2.3.4. Training and Evaluation Process

Several variations of the three different deep learning methods were applied in order to counter class imbalance and investigate the influence of convolutional layer parameters in the model's performance. Both ResNet and DenseNet models were trained using only original and a combination of original and augmented images through clockwise, anti-clockwise rotation (±5°) and contrast ($\times 1.5$) and brightness ($+50$ in each pixel value) enhancement in both original and rotated images using the *cv2.convertScaleAbs* function of OpenCV image processing library. In order to enhance the attention of those two models on the under-represented progression group, class weights were calculated using the *sklearn.class_weight* function of the Scikit Learn Python library for training with the original data. Furthermore, ResNet and DenseNet were trained with the use original and augmented data and the implementation of L_2 norm regularization term in each convolution layer, resulting in a total of five different models for each of the two architectures (trained on original data, original data + L_2 norm, original data + L_2 norm + class weights, augmented data, augmented data + L_2 norm). In the case of CVAE method, the only augmentation method was the oversampling (creating duplicates) of the progression group due to the structure of the model that required the presence of the same number of subjects from each class. The addition of L_2 norm regularization was also implemented in this method, leading the creation of two main CVAE models, with and without regularization.

In order to incorporate patient's clinical data, a Logistic Regression classifier was implemented. Regarding ResNet and DenseNet models, their output probability value was combined in a vector with patient's age, gender and BMI values, while in the case of CVAE the learned feature vector was concatenated with clinical information for the creation of the classifier's input.

The initial control and progression groups were split into 70/15/15 ratio resulting into 425 training, 77 validation and 89 testing 3D IW-TSE images, and data augmentation was applied only in the training dataset. During the training process, the Adam optimizer method was implemented for each of the three deep learning architectures. Several different values of model hyper-parameters were tested. For ResNet and DenseNet those hyper-parameters were the input batch size and the Adam's learning rate, while for the CVAE the values of epoch number, discriminative penalty term, latent space size, batch size and Adam's learning rate were changed.

The performance of each deep learning method along with the Logistic Classifier on the testing dataset was assessed using AUC and PR-AUC metrics.

3 Results

The outcomes of the previously described deep learning architectures are presented in this chapter. These include the final number of subjects in control and case groups, the training and testing results of the U-net for the segmentation of the 3D DESS MRI scans, the creation of IW-TSE masks through image registration and the tables and figures of the detection algorithms' performance (ResNet-50, DenseNet-121, CVAE).

3.1. Control and Case groups

After a thorough search of the files available at OAI that contained MOAKS, WORMS and KL grading scores of OA features, the basic information regarding the control and case groups are shown in table 3.1:

OA group	Control	Case
# subjects	434 (73%)	159 (27%)

Table 3.1: Control and case groups.
Control group: number of participants that did not develop knee OA after 24 months.
Case group: number of participants that developed knee OA after 24 months.

3.2. U-Net

In the current section the DESS sequences segmentation performance with different values in the model's hyper-parameters along with the DESS-to-IW-TSE registration outcomes are going to be presented. Representative images of the created and registered masks are also included.

3.2.1. Segmentation

507 annotated 2D sagittal DESS MRI sequences were used in order to train the U-Net for image segmentation. Regarding the model's hyper-parameters, the number of epochs was kept constant (5) and three different batch sizes were tested (4,8,16). In table 3.2 the mean values for all classes for the training performance metrics with different batch size are presented.

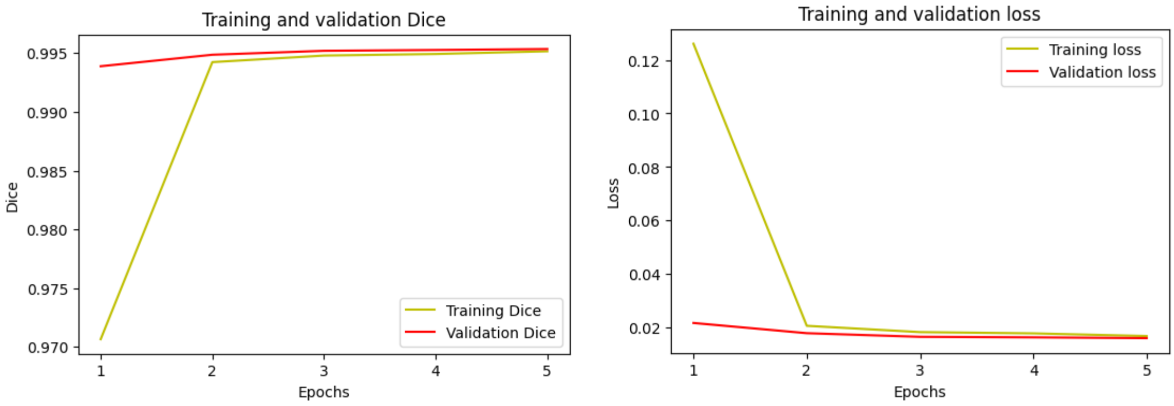
Cumulative Results

Batch Size	Dice Coefficient	Loss Function	Accuracy	mean IoU
4	0.99509	0.01711	0.99334	0.88835
8	0.99539	0.01607	0.99371	0.89311
16	0.99542	0.01544	0.99391	0.90538

Table 3.2: Testing performance metrics for all different batch sizes (4,8,16)

Based on table 3.2, the U-Net achieved higher performance metrics with batch size 16, since not only the Dice coefficient was 0.99542 but also the mean Intersection of Union exceeded 0.90. Thus, these configurations were chosen in order to create the bone and cartilage labelled images for the 594 DESS MRI scans of the control and case groups. In the figure 3.1 the training metrics of the U-Net with batch size 16 and in table 3.3 for tibial and femoral classes of the same batch size are presented.

Batch size 16



(a) Dice Similarity Coefficient

(b) Loss Value

Figure 3.1: Training and validation graphs for
Dice similarity coefficient and loss function values for batch size 16

	Dice Coefficient	Intersection of Union
Tibia	0.9846893	0.9698405
Femur	0.9870652	0.9744608

Table 3.3: Dice Similarity Coefficient and Intersection of Union values
for tibial and femoral labels for batch size 16

Segmentation Results

Figures 3.2, 3.3, 3.4 and 3.5 show some of the segmentation outputs of the 594 DESS MRI scans of both the control and progression groups are presented.

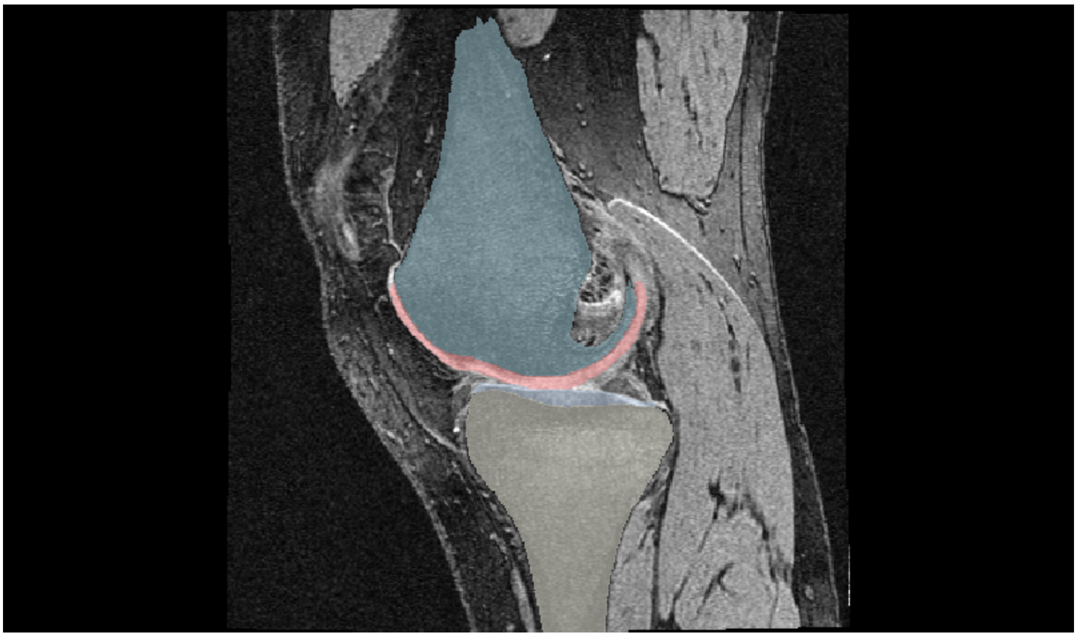


Figure 3.2: Sagittal view of right knee of control group patient 9725978, slice 90
Superimposed original and segmented DESS MRI

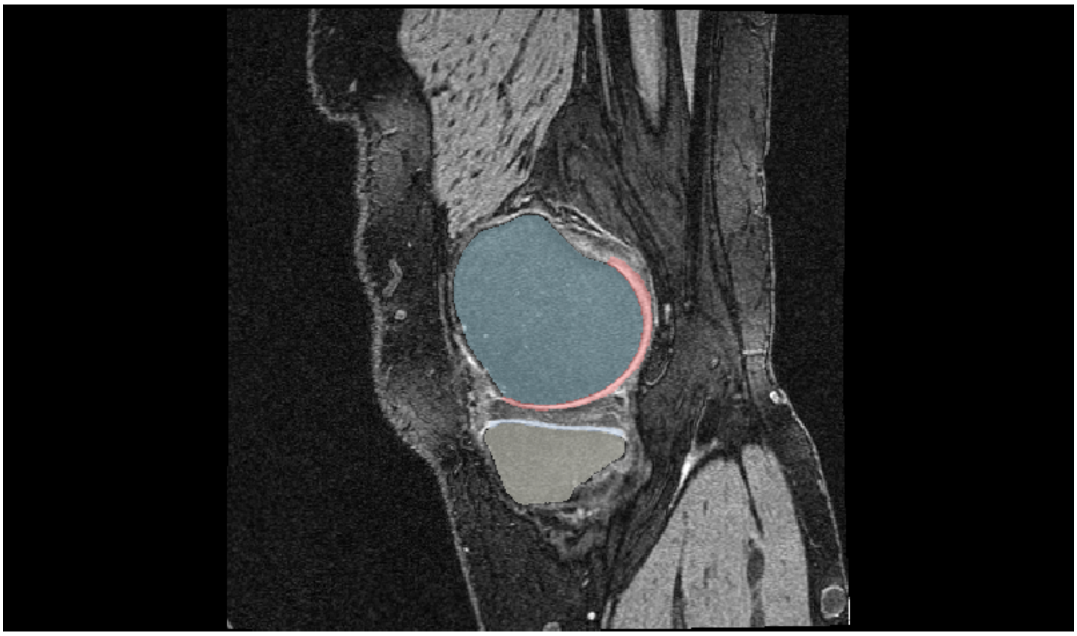


Figure 3.3: Sagittal view of left knee of control group patient 9412037, slice 41
Superimposed original and segmented DESS MRI

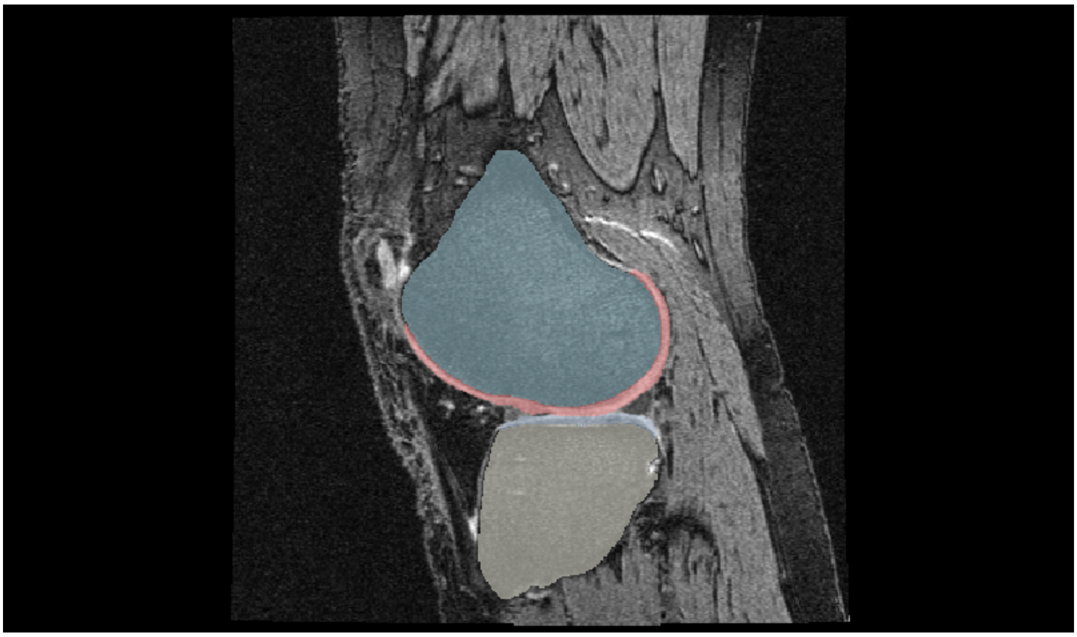


Figure 3.4: Sagittal view of left knee of progress group patient 9029791, slice 116
Superimposed original and segmented DESS MRI

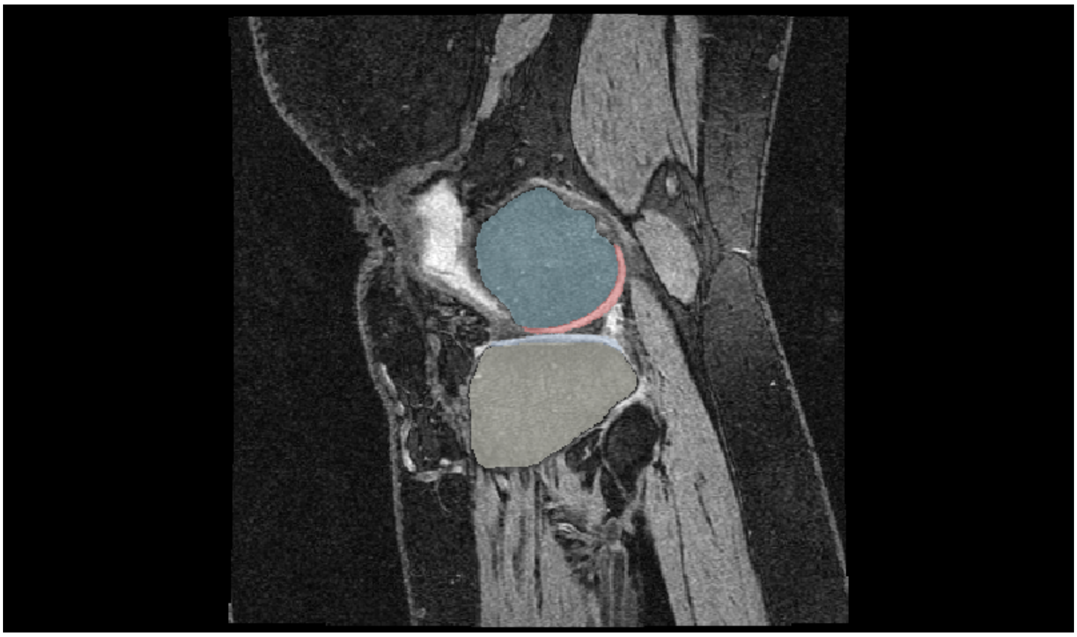


Figure 3.5: Sagittal view of right knee of progress group patient 9878804, slice 40
Superimposed original and segmented DESS MRI

3.2.2. Registration

In figure 3.6, the DESS-to-IW-TSE registration along with the created labelled IW-TSE MRI scans are presented. In figure 3.7 the alignment of the output IW-TSE mask over the original IW-TSE MRI sequence of one subject is shown as an example of the visual inspection of the registration process result.

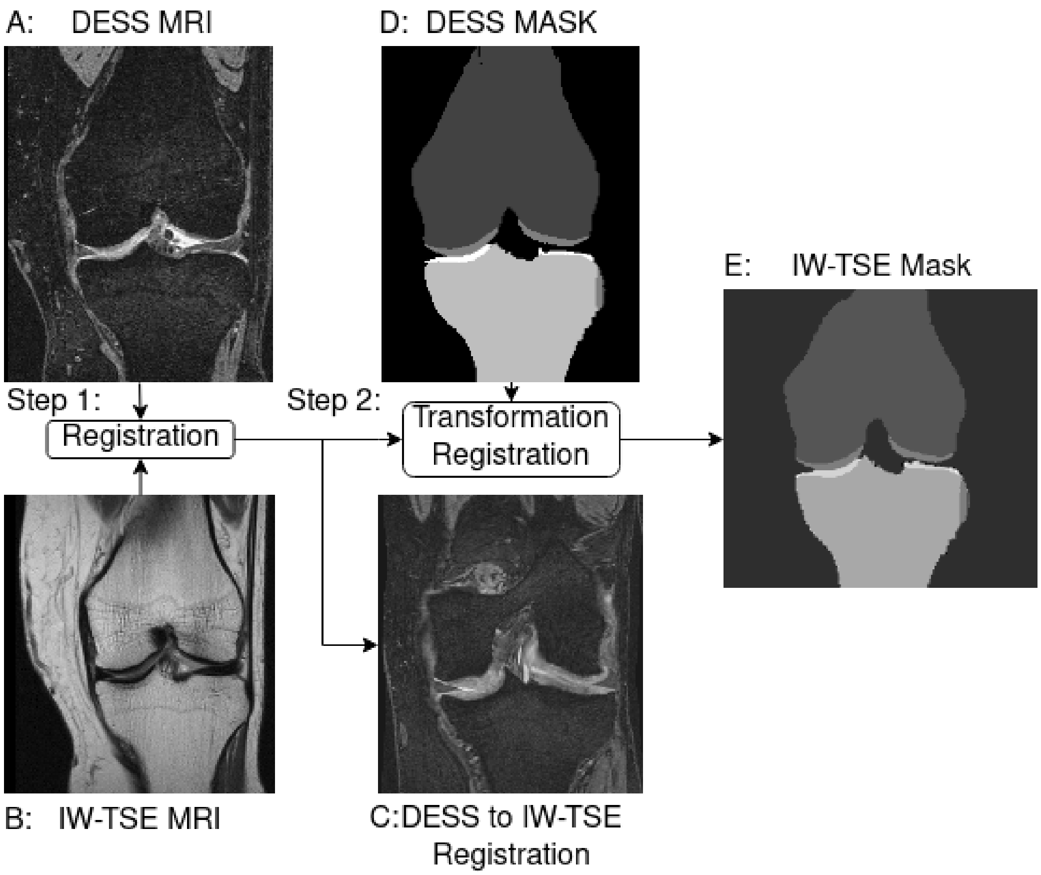
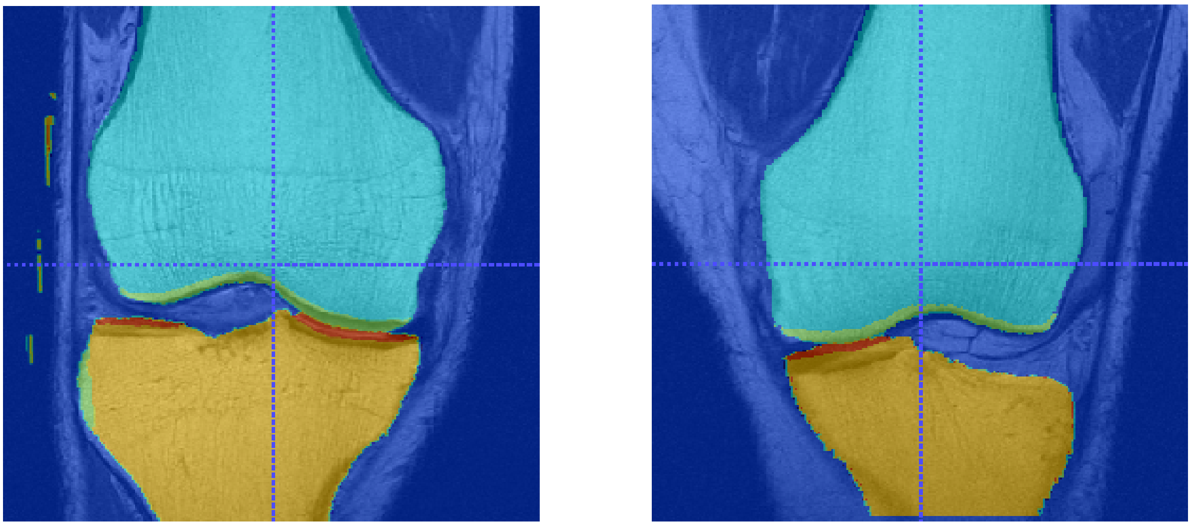


Figure 3.6: Registration pipeline of left knee for patient 9050299.
Step 1: DESS (A) to IW-TSE (B) registration with output DESS-to-IW-TSE image C.
Step 2: DESS Mask(D) to DESS-to-IW-TSE(C) registration.
Output the desired IW-TSE Mask (E)

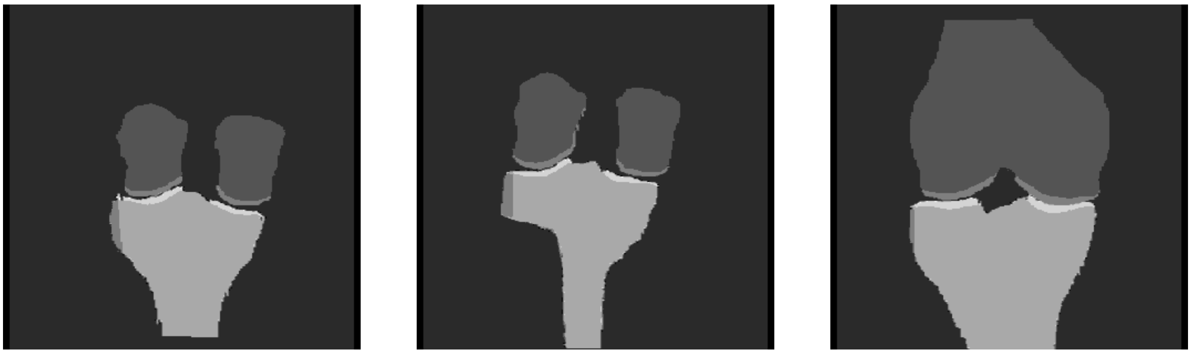


(a) 9003113-R

(b) 9890414-L

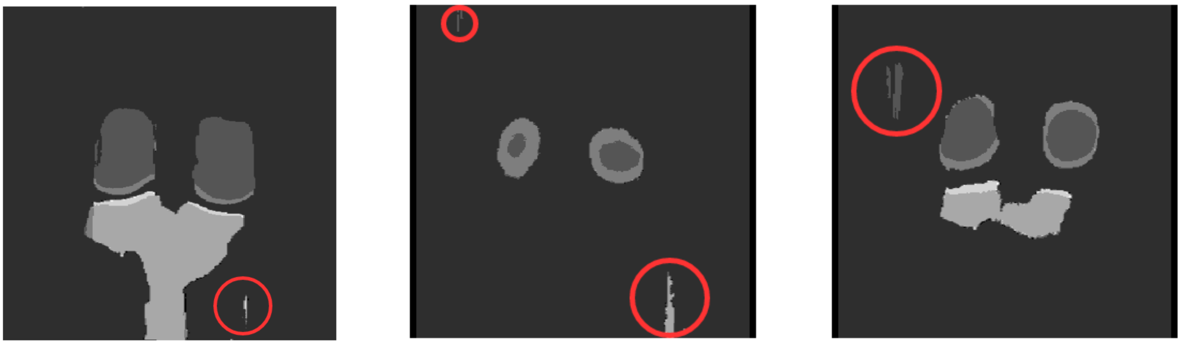
Figure 3.7: Alignment of IW-TSE MRIs and their corresponded IW-TSE masks.

In the figures 3.8 and 3.9 several examples of the registration output IW-TSE masks are presented. In several cases some background areas were labelled as bone or cartilage regions, which led to the implementation of Canny Edge Detector for the purpose of dismissing those misclassified pixels for minimum and maximum bone coordinates search. After trials and errors with the Canny Edge Detection parameters, the values that led to dismissing the incorrect pixels were variance $\sigma = 8$, for the tibial bone low threshold $= 0.90$ and high threshold $= 0.95$, and for femoral bone low threshold $= 0.37$ and high threshold $= 0.39$.



(a) 9609732-R **(b)** 9866291-R **(c)** 9956822-R

Figure 3.8: Examples IW-TSE masks after the registration process *without* misclassified background pixels.



(a) 9248710-R **(b)** 9282888-R **(c)** 9830048-R

Figure 3.9: Examples IW-TSE masks after the registration process *with* misclassified background pixels.

3.2.3. ROI Extraction

The next step after the registration and the creation of IW-TSE MRI masks, was the application of whole dataset search in order to find the minimum and maximum bone coordinates. The results of this search are presented in the table 3.4

Based on the minimum and maximum bone coordinates, the region of interest around the knee joint would have width $=$ 306 pixels and height $=$ 155 pixels, with starting point (33,133). After a visual inspection of the created rectangular it was decided to increase its dimensions and change the starting point to (33,93) in order to include as much femoral bone as possible.

Coords	Femur Xmin	Femur Xmax	Tibia Xmin	Tibia Xmax
X	33	339	43	320
Y	133	209	197	288

Table 3.4: Minimum & Maximum Bone Coordinates
of both control and case groups

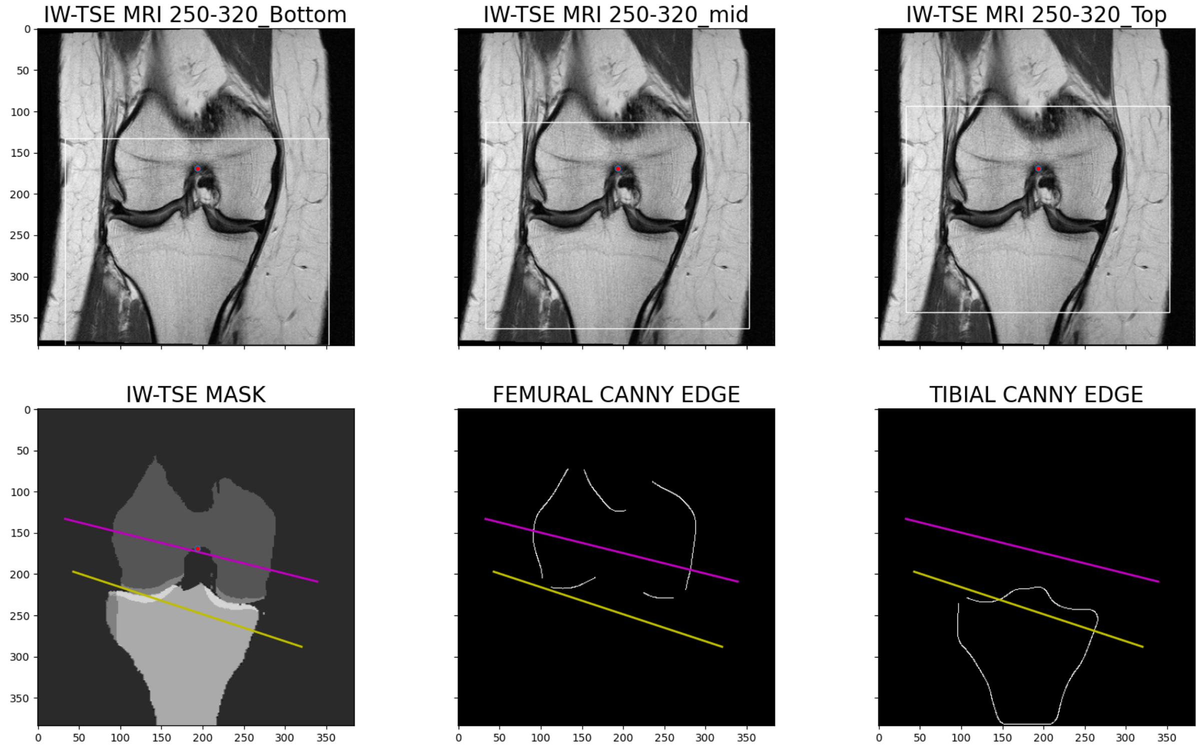
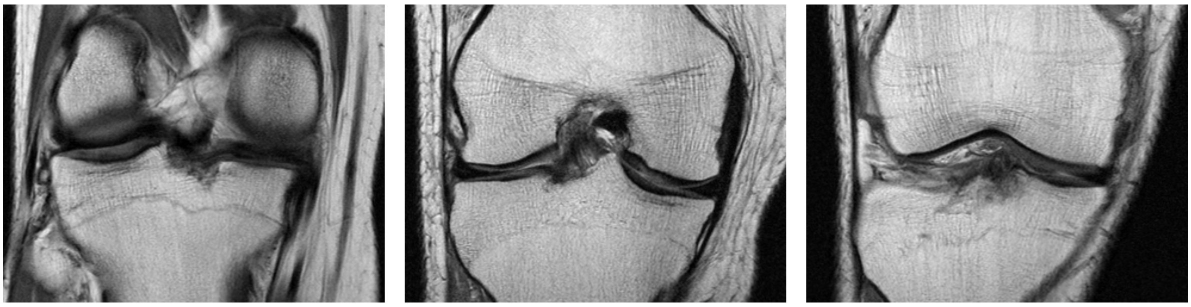


Figure 3.10: Different starting points of ROI (top). Minimum and maximum of tibial and femoral bones with femoral and tibial canny edge detection (bottom). Yellow line connects minimum and maximum tibial coordinates. Purple line connects minimum and maximum femoral coordinates

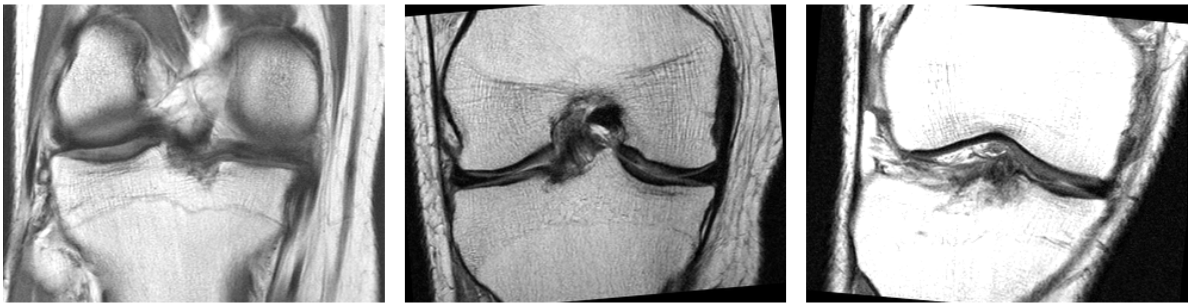


(a) 9457264-R

(b) 9004184-R

(c) 9885588-R

Figure 3.11: Examples of final ROIs of 250×320 pixel size.



(a) 9457264-R

(b) 9004184-R

(c) 9885588-R

Figure 3.12: Augmentation examples of final ROIs of 250×320 pixel size.
(a) brightness adjustment, (b) anti-clockwise rotation,
(c) clockwise rotation and contrast adjustment

3.3. Residual Network (ResNet)

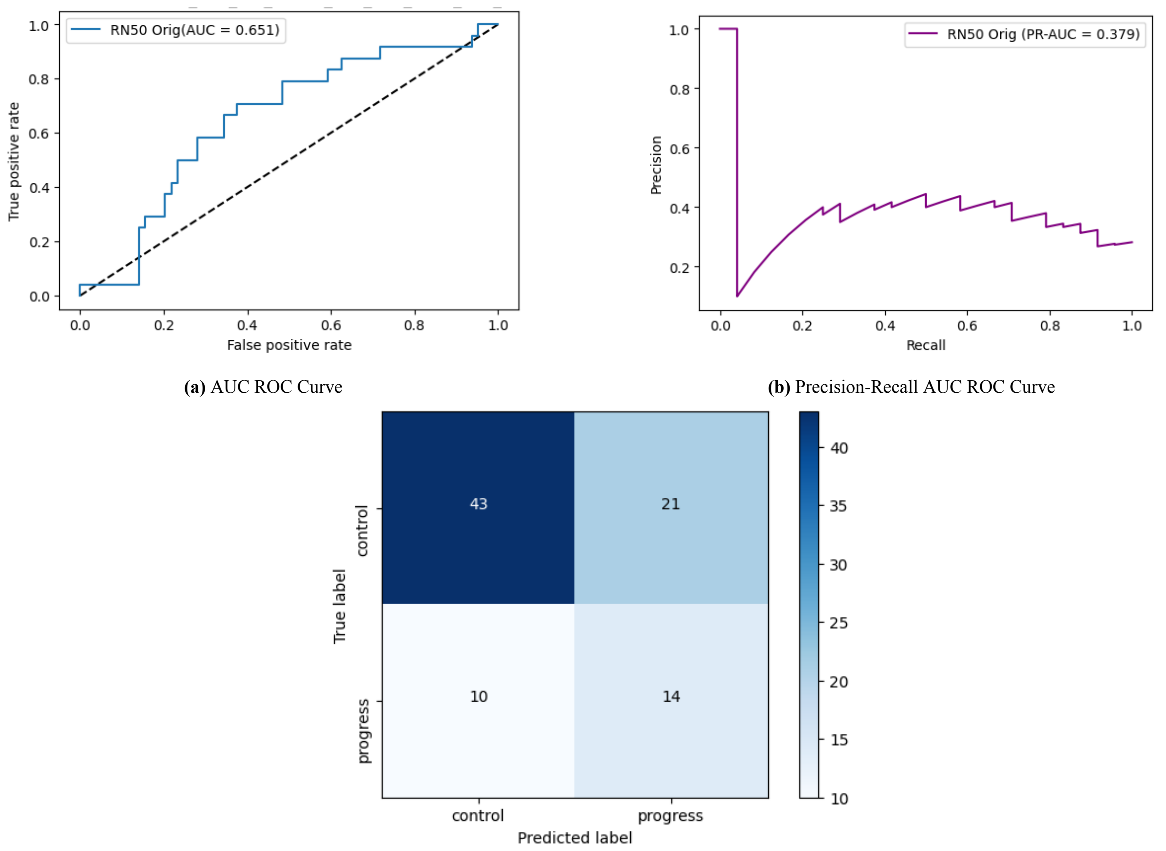
The cumulative performance metrics of ResNet-50 are presented in the Appendix, with changed values regarding the batch size and the Adam's optimizer learning rate. Table 3.5 shows the AUC and the PR-AUC of the highest performing ResNet-50 models.

			ResNet-50		ResNet-50 + Clinical Data	
	Batch Size	Adam's Learning Rate	**AUC**	**PR-AUC**	**AUC**	**PR-AUC**
orig+reg	4	0.01	0.6042	0.3194	0.6406	0.3613
orig+reg	8	0.01	0.5983	0.3115	0.6511	0.3786
orig+aug+reg	8	0.01	0.5846	0.3912	0.6458	0.3669
orig+aug+reg	4	0.001	0.5775	0.3752	0.6458	0.3633

Table 3.5: ResNet-50 AUC & PR-AUC values for epochs number= 15
batch size= 8 & 4 and Adam's optimizer learning rate= 0.001 & 0.01 .
orig+reg: trained on original data with kernel regularization
orig+aug+reg: trained on original and augmented data with kernel regularization

It is important to state that the Logistic Regression Classifier, when trained only with clinical data (age, BMI, gender), achieved an AUC of 0.6399 and a PR-AUC of 0.3643. Regarding the performance of the ResNet-50 with batch size 4, the highest AUC and PR-AUC without the incorporation of patients' data in the classifier was 0.6042 and 0.3194, respectively, for training data without any augmentation technique and the use of kernel regularization terms in its layers. When the output probabilities of the ResNet-50 with batch size 4 were combined with age, BMI and gender for the training and testing of the logistic regression classifier, the highest AUC was 0.6458 with PR-AUC at 0.3633. This values refer to the ResNet-50 that used augmented training data and kernel regularization terms.

When ResNet-50 was trained with batch size values equal to 8, the highest AUC value, 0.5983, with corresponding PR-AUC value of 0.3115, were reached by the model version trained on original data and kernel regularization. The same trained model variation reached the highest values in both AUC (0.6511) and PR-AUC (0.3786) when its output was combined clinical data to be used as input to the classifier. Figure 3.13 presents the AUC, PR-AUC ROC Curves and the confusion matrix for the best performing classifier that combines the output of the ResNet-50 model and clinical data .



(a) AUC ROC Curve

(b) Precision-Recall AUC ROC Curve

(c) Confusion matrix

Figure 3.13: ResNet-50 in combination with Logistic Regression classifier,
epochs: 15, batch size: 8, Adam's learning rate: 0.01,
use of kernel regularization terms, trained on original dataset

3.4. Densely Connected Convolutional Network (DenseNet)

Regarding the performance of the DenseNet-121, the model's highest performance metrics with different hyper-parameters' values and inputs are shown in table 3.6.

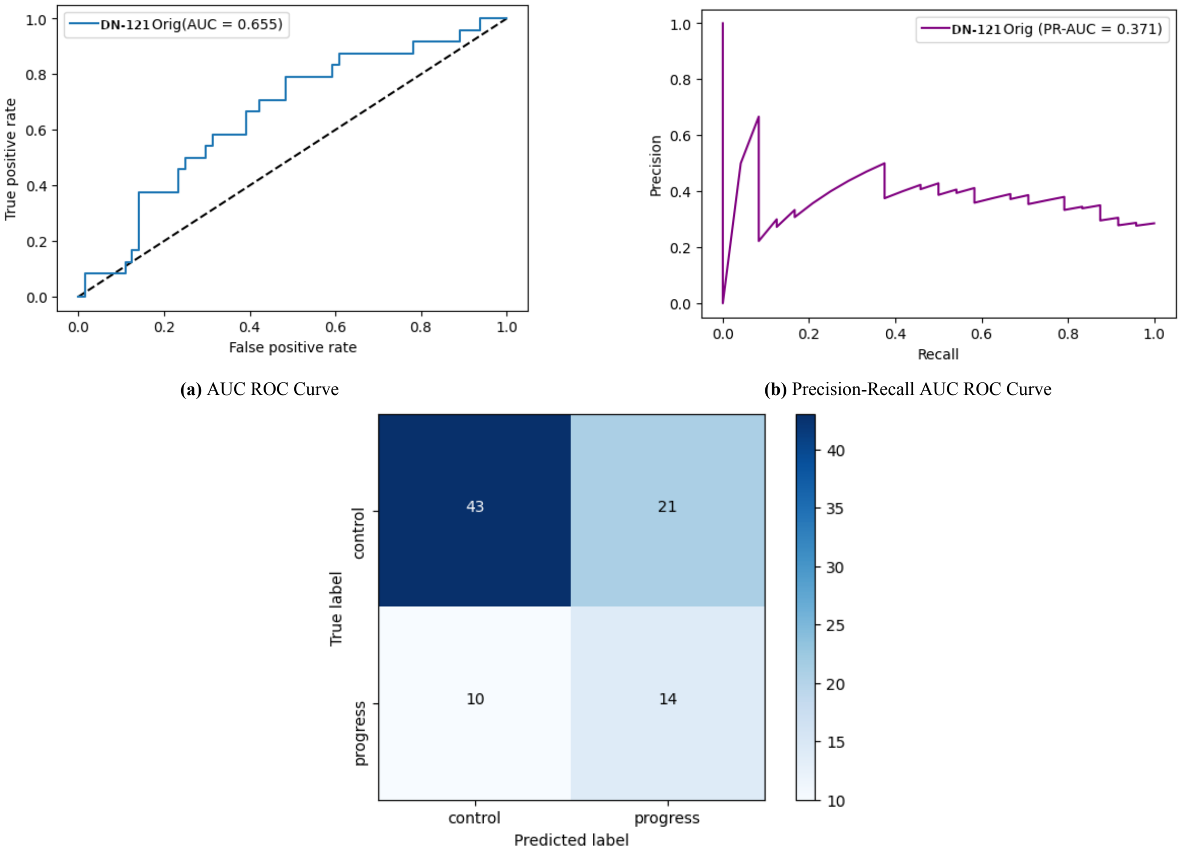
			DenseNet-121		DenseNet-121 + Clinical Data	
	Batch Size	Adam's Learning Rate	**AUC**	**PR-AUC**	**AUC**	**PR-AUC**
orig	8	0.01	0.6269	0.3711	0.6406	0.3653
orig+reg	4	0.01	0.5618	0.2945	0.6556	0.3714
orig+reg+bw	4	0.01	0.5521	0.2797	0.6198	0.3217
orig+aug	8	0.001	0.5358	0.3641	0.6543	0.3623
orig+reg	4	0.001	0.5273	0.2728	0.6426	0.3386

Table 3.6: DenseNet-121 AUC & PR-AUC values for epochs number= 15,
batch size= 8 & 4 and Adam's optimizer learning rate= 0.001 & 0.01 .
orig: trained on original data, orig_reg: trained on original data with kernel regularization
orig+reg+bw: trained in original data with kernel regularization and balanced weights
orig+aug: trained on augmented data

Regarding the performance of the DenseNet-121 with batch size 4, the model that was trained with original data and the use of kernel regularization terms achieved the highest AUC value for both cases, with and without the incorporation of clinical data in the logistic regression classifier. When only the model outputs were counted, the AUC and PR-AUC values were 0.5618 and 0.2945 respectively, and the combination of model outputs and age, BMI and gender information reached an AUC of 0.6556 and a PR-AUC of 0.3714.

When DenseNet-121 used batch size 8, the highest AUC regarding the model's output was reached by the version trained with original data, with an AUC of 0.6269 and a PR-AUC of 0.3711. When clinical data were combined with the ResNet probability values for the classifier, the AUC and the PR-AUC were 0.6543 and 0.3623, respectively.

AUC ROC and PR-AUC ROC Curves along with the confusion matrix of the best performing DenseNet + Logistic Regression classifier are presented in figures 3.14.



(a) AUC ROC Curve

(b) Precision-Recall AUC ROC Curve

(c) Confusion matrix

Figure 3.14: DenseNet-121 in combination with Logistic Regression classifier, epochs: 15, batch size: 4, Adam's learning rate: 0.01, use of kernel regularization terms, trained on original dataset

3.5. Convolutional Variational Autoencoder (CVAE)

Several models of Convolutional Variational Autoencoder were tested for the task of early detection of knee OA through MRI scans, due to the increased number of hyper-parameters that are included in this deep learning algorithms. Apart from the number of epochs (25, 50, 100, 200), batch size (4,8) and Adam's learning rate(0.001, 0.01), different values of the discriminative penalty (0.0005, 0.001, 0.01) in the loss function and the size of the latent feature space (100, 500, 1000) were tested. In total 80 CVAE models were tested. In tables 3.7, 3.8 CVAE algorithms with the highest AUC values are presented.

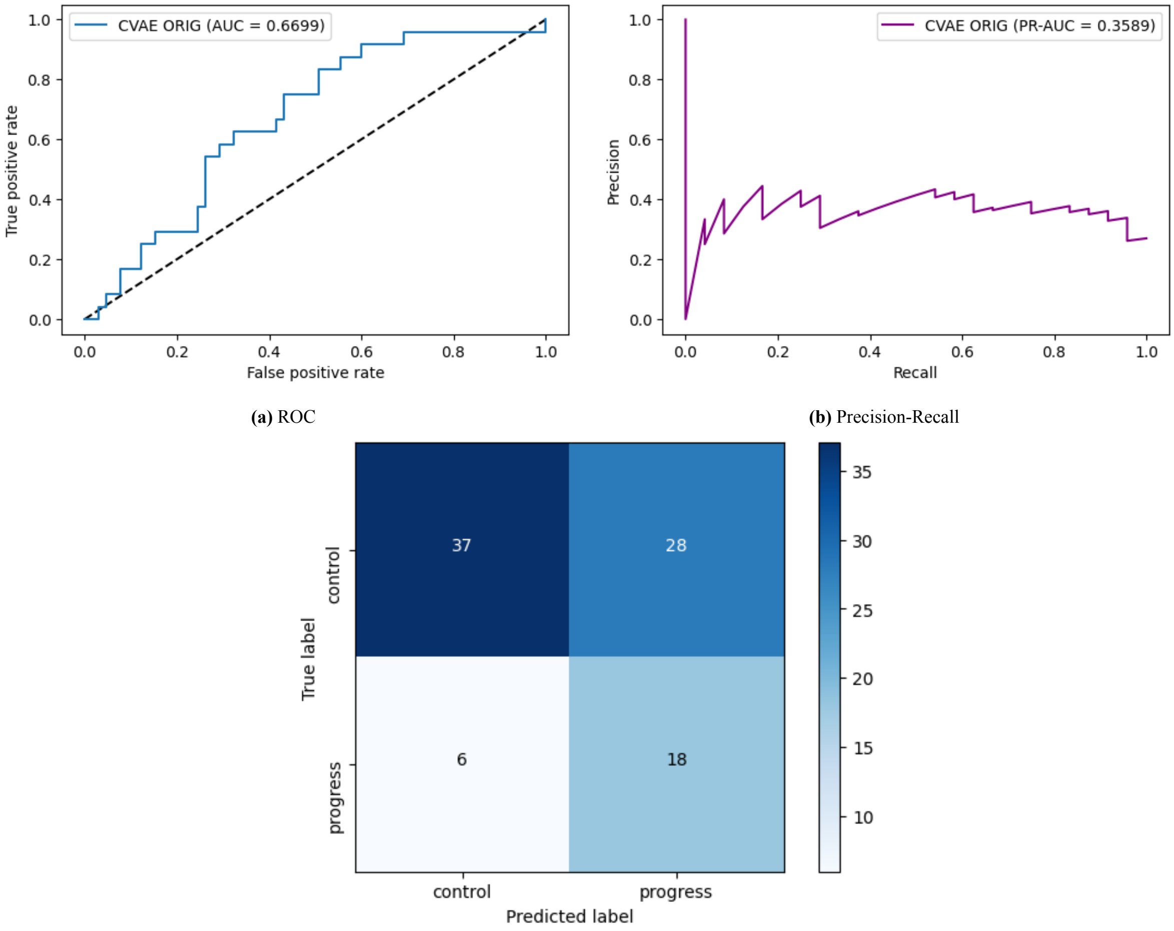
Epochs: 100	Without Kernel Regularization			
	CVAE and Logistic Regression		CVAE in combination with Clinical Data in Logistic Regression	
	AUC	**PR-AUC**	**AUC**	**PR-AUC**
BS=4, LR=0.01, DiscPen=0.01, CD=1000	0.6689	0.3463	0.6699	0.3589
BS=4, LR=0.001, DiscPen=0.01, CD=1000	0.6468	0.3623	0.6564	0.3651
BS=4, LR=0.01, DiscPen=0.001, CD=1000	0.6442	0.3625	0.6468	0.3654
BS=4, LR=0.01, DiscPen=0.001, CD=100	0.6006	0.3164	0.6628	0.3523

Table 3.7: CVAE in combination with Logistic Regression classifier, without kernel regularization terms, trained on balanced dataset (BS=batch size, LR= Adam's learning rate, DiscPen= discriminative penalty, CD=latent space dimensions.)

Epochs: 100	With Kernel Regularization			
	CVAE and Logistic Regression		CVAE in combination with Clinical Data in Logistic Regression	
	AUC	**PR-AUC**	**AUC**	**PR-AUC**
BS=8, LR=0.001, DiscPen=0.01, CD=1000	0.6064	0.4038	0.6218	0.4372
BS=4, LR=0.001, DiscPen=0.01, CD=500	0.6006	0.4161	0.6141	0.4101
BS=8, LR=0.01, DiscPen=0.001, CD=100	0.5705	0.2978	0.6276	0.3384
BS=4, LR=0.001, DiscPen=0.001, CD=100	0.5647	0.2904	0.6288	0.3291

Table 3.8: CVAE in combination with Logistic Regression classifier, with kernel regularization terms, trained on balanced dataset (BS=batch size, LR= Adam's learning rate, DiscPen= discriminative penalty, CD=latent space dimensions.)

For the above tables, the best performing CVAE model has batch size = 4, learning rate = 0.01, discriminative penalty = 0.01 and latent space dimension = 1000, for both logistic regression inputs. With only the CVAE learned code as input the logistic regression achieved an AUC of 0.6689 and a PR-AUC of 0.3463, and with the addition of clinical data the AUC was 0.6699 and the PR-AUC 0.3589.



(a) ROC

(b) Precision-Recall

(c) Confusion matrix

Figure 3.15: CVAE and Clinical Data with Logistic Regression classifier, epochs: 100, batch size: 4, learning rate: 0.01, discriminative penalty:0.01, latent space dimensions: 1000

Table 3.9 presents the highest AUC values which were achieved for each algorithm combination with the Logistic Regression Classifier.

	Parameters							Metrics	
	Trained Data	Kernel Reg.	Epochs	Batch Size	Adam's Learning Rate	Disc. Pen.	Code Dim.	**AUC**	**PR-AUC**
RN-50	Orig.	No	15	8	0.01	-	-	0.6511	0.3613
DN-121	Orig.	Yes	15	4	0.01	-	-	0.6556	0.3714
CVAE	Orig.+Aug.	No	100	4	0.01	0.01	1000	0.6699	0.3589

Table 3.9: Best detection models & clinical data as Logistic Regression Classifier's input
RN-50: ResNet-50, DN-121: DenseNet-121, CVAE: Convolutional Variational Autoencoder
Kernel Reg.: L_2 regularization term, Disc.Pen.: Discriminative Penalty

4

Discussion

4.1. Discussion

The main aim of the current master thesis project was to examine the influence of deep learning-based MRI features in the early detection of knee OA progression. Several different steps were involved in order to investigate the main research question of this thesis. Initially, the OAI studies that include subjects with knee MRI scans were acquired and the features that define the presence of OA were searched. Due to missing osteophyte measurements and feature scores in every time point, a modified OA definition was applied that included BML, cartilage thickness loss and meniscal degradation scores from MRI-based grading systems and osteophyte scores from X-ray based system, leading to the creation of control and case groups of subject that developed OA within the first 24 months from their initial visit. The next step involved the development of a semi-automated method for the extraction of the knee joint region, that included the segmentation through a U-Net of DESS MRI sequences, their registration to IW-TSE scans and the creation of the rectangular knee region based on tibial and femoral bone coordinates. This process resulted in the construction of the final 3D IW-TSE images, which were given as input to three different deep learning algorithms (ResNet, DenseNet, CVAE) for the task of early detection of knee OA. The output of these DL methods was combined with clinical data (age, gender, BMI) and given as input to a logistic regression classifier in order to additionally examine the effect of non-imaging features to early OA progression diagnosis. The performance of segmentation method (U-Net) was measured using the Dice similarity coefficient and the IoU and the three different early detection algorithms' performance was measured using the AUC and PR-AUC metrics. The highest AUC value (0.6689) was achieved by the CVAE method with batch size 4, Adam's learning rate 0.01, discriminative penalty 0.01 and feature vector of 1000 elements, when the algorithm's output was combined with clinical information as input to the logistic regression classifier. A significant remark about the results was that the combination of patients' data with the output probability value of the deep learning algorithms in order to construct the input vector to the logistic classifier yielded higher AUC and PR-AUC values that the detection algorithms alone.

The developed semi-automated ROI creation method involved the segmentation of DESS MRI scans, the registration of DESS to IW-TSE images and the search of minimum and maximum tibial and femoral bone coordinates in the newly created IW-TSE masks. The applied U-Net algorithm achieved the highest Dice similarity coefficient value and mean IoU when trained with batch size 16, 0.99542 and 0.90538 respectively, as was reported in table 3.2, similar to the performance of the U-Net developed by Ambellan [45]. For the same batch size the IoU values for tibial and femoral bones were 0.9698405 and 0.9744608 respectively, while the Dice coefficient similarity values were 0.9846893 and 0.9870652 respectively. These performance metrics showed that the implementation of U-Net for sagittal DESS segmentation could yield sufficient bone labelled images, which could assist in the construction of knee joint masks in each plane (coronal, axial, sagittal). Regarding the registration of DESS masks to IW-TSE MRI sequences with the goal of IW-TSE bone and cartilage label formation, the discarding of one MRI scan due to errors along with the visual inspection of the created masks could suggest room for improvement in the direction of multimodel image registration, however this was out of the scope of this thesis project.

Concerning the trained deep learning architectures that were tested for the detection of early progression of knee OA through IW-TSE MRI sequences, the AUC values that were found for the datasets incorporated in this project were lower than the majority of the existing literature which for both X-ray and MRI modalities achieved AUC greater than 0.80, as presented in section 1.4 of the introduction. This holds for both outputs, only the use of DL methods probability output and their combination with patients' clinical data (age, gender, BMI). The addition of these variables that are associated with OA presence in knee joints increased the AUC and PR-AUC performance metrics for every trained model (ResNet-50, DenseNet-121, CVAE) that was tested. The highest AUC value for ResNet-50 with only 3D MRI scans as input was 0.6042 (batch size=4, Adam's learning rate=0.01), while when it's probability value was combined with patients' information the AUC reached 0.6541 (batch size=8, Adam's learning rate=0.01). Similarly for DenseNet-121 the best performing model without clinical information had an AUC of 0.6269 (batch size=8, Adam's learning rate=0.01), while with the combination of age, gender and BMI an AUC of 0.6556 (batch size=8, Adam's learning rate=0.01). The increase of the highest AUC value for the tested CVAE when additional data were combined as input to the logistic regression classifier was smaller, 0.6689 for CVAE with only image input and 0.6699 for CVAE probability value and clinical data as input to the classifier.

Moreover, regarding the two methods with similarities in their architecture design (ResNet and DenseNet), there were not significant differences in their detection measurements when only the model's outputs were used, since there was the same number of model' versions with AUC under 0.5 (six for both ResNet and DenseNet). The effect of the clinical data combination with model's outputs to form the logistic regression input was greater in DenseNet-121 since more models achieved an AUC larger than 0.6 (18 vs 12). One reason for this could be the extensive preservation of information from previous layer outputs applied in DenseNet models, which concatenate feature maps from a larger number of layers inside the dense blocks.

The tested CVAE algorithm was a variation of a similar method published by Nasser[23], where an autoencoder with only dense layers was applied for the task of early detection of knee OA from X-rays. The version of the autoencoder developed in this project included 3D convolutional layers and the reparameterization of the encoded latent space in order to extract spatial and more smooth features from the 3D input MRI scans. The representation code of the CVAE was used as input to the logistic regression classifier alone and in combination with clinical data, and showed adequate performance metrics when the number of epochs during the training process was 100, reaching AUC values between 0.55 and 0.6. With the differentiation of the latent vector size, from 100 to 500 and 1000, the performance metrics were improved and reached their highest value for code size 1000, for both logistic regression inputs (code vector alone, combination of code and clinical data), showing a similar behavior as the autoencoder developed by Nasser[23], where the highest accuracy (0.81) was achieved when code size was 1000 and discriminative penalty 0.01.

The CVAE model with batch size 4, epoch number 1000, Adam's learning rate 0.01, latent space dimension 1000 and without the use of L_2 regularization term achieved the highest AUC values when coupled with the logistic regression classifier for both inputs, reaching an AUC of 0.6689 with only code input and an AUC of 0.6699 with the combination of code and clinical information as input. The latter AUC value was the highest among the 80 CVAE, 12 ResNet and 12 DenseNet models trained for the task of early detection of knee OA from MRI scans. Comparing CVAE with the other two tested methods (ResNet-50, DenseNet-121) showed that the CVAE model was able to distinguish more informative features from the 3D input IW-TSE MRI scans. This higher AUC value of the CVAE method could also indicate that the autoencoder's texture analysis can extract meaningfull patterns between early OA patients and healthy subjects, when compared to ResNet and DenseNet algorithms, which perform an analysis based on the shape of the input images.

Several factors might limit the performance of the developed methods and could explain their lower metrics when compared to the existing literature as mentioned in section 1.4. The method applied for the construction of the control and progression groups might have hindered the developed DL algorithms' ability to detect OA progression, since the distinction between case and healthy subjects was based solely on their KL grades and not on a MRI based scoring method. The incorporation of a modified MOAKS grading scale for the definition of OA in knee MRI due to the lack of osteophyte measurements also influenced the number of subjects constructing both control and case groups, and especially the latter, since MR imaging modality has higher sensitivity and specificity in osteophyte detection. Furthermore, another aspect that might reduced the models' performance was the number of clinical variables used as input, since previous publications did not only include age, gender and BMI but also clinical data such as knee injury and surgery history or an expert defined KL grade. Moreover, the developed algorithm's were trained using random initialization of their weights, while the majority of the published algorithms incorporated pre-trained weights, which can improve their AUC values on knee OA detection and prognosis.

4.2. Future research

Some suggestions and recommendations that could validate and expand the knowledge obtained in the current master thesis project regarding the influence of deep learning-based MRI data on early detection of knee OA are:

1. The influence of smaller areas of the knee joint on the detection of OA progression could be examined via the incorporation of heat maps in the models' outputs or via the use of smaller ROIs (medial and lateral knee side).
2. In this project the performance of a Convolutional Variational Autoencoder with a discriminative term in its loss function was investigated. Several different implementations of 3D Autoencoders can be examined, not only with convolutional but also with an addition of dense layers [23].
3. In terms of the ResNet and DenseNet models' structure, the recommendation would be to test their deeper variants (ResNet-101,-152, DensNet-169,-201,-264), since these types are most commonly used in the existing publications [21, 26].
4. The effect of pre-trained weights in the tested detection algorithms (ResNet-50, DenseNet-121, CVAE) could be investigated since the majority of the published research applied transfer learning for the tested DL algorithms[16–18, 23, 27].
5. A different ROI extraction algorithm could be applied such as Regional Proposal Network (RPN)[49] and YOLO[50], since the developed pipeline incorporated many steps and procedures (bone segmentation, multi-modal image registration, semi-automatic ROI definition) which can introduce errors and require more time and computational resources.
6. Different MRI sequences can be used to train the same deep learning algorithms in order to examine their influence.

5

Conclusion

For this master thesis project the influence of deep learning-based MRI features in the early detection of knee osteoarthritis progression was examined, which included the investigation of three different deep learning architectures along with a semi-automated construction of a segmentation method for the knee joint extraction.

The results of the region-of-interest creation showed promising outcomes for the application of a U-Net algorithm for the bone and cartilage tissue labelling, reaching high accuracy and Dice Similarity Coefficient metric values.

Existing publicly available deep learning models have focused mainly on the OA severity detection through X-ray and MRI scans, reaching AUC values of 0.90 and 0.82 for modality, with a few publications examining the early detection of knee osteoarthritis progression only through X-ray. The maximum AUC value when X-ray scans and clinical data from OAI have been processed for OA prediction progression is 0.81 .

In this study, the results showed that different deep learning algorithms have similar performance metrics when 3D MRI scans are being processed since ResNet, DenseNet and CVAE had close AUC and PR-AUC values. Furthermore, the combination of model predictive outcomes and clinical data as input for the Logistic Regression classifier increased their performance, showing the influence of OA-related variables such as age, gender and BMI to the detection of knee OA.

Previous studies have already shown the ability of deep learning methods to detect and classify knee osteoarthritis through mainly X-rays scans. This study contributes in the same field, examining the effect of MRI scans for the early progression of OA. Deep learning models with the incorporation of MRI sequences as inputs although they do not reach the same level of performance metrics have the potential to influence in a more informative and elaborate way the early detection of knee osteoarthritis.

Bibliography

[1] Nigel Arden and Michael C Nevitt. “Osteoarthritis: epidemiology”. In: *Best practice & research Clinical rheumatology* 20.1 (2006), pp. 3–25.

[2] Marita Cross et al. “The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study”. In: *Annals of the rheumatic diseases* 73.7 (2014), pp. 1323–1330.

[3] Sion Glyn-Jones et al. “Osteoarthritis”. In: *The Lancet* 386.9991 (2015), pp. 376–387.

[4] David T Felson. “Osteoarthritis of the knee”. In: *New England Journal of Medicine* 354.8 (2006), pp. 841–848.

[5] Anuradha Vashishtha and Anuja kumar Acharya. “An overview of medical imaging techniques for knee osteoarthritis disease”. In: *Biomedical and Pharmacology Journal* 14.2 (2021), pp. 903–919.

[6] Mark D Kohn, Adam A Sassoon, and Navin D Fernando. “Classifications in brief: Kellgren-Lawrence classification of osteoarthritis”. In: *Clinical Orthopaedics and Related Research®* 474.8 (2016), pp. 1886–1893.

[7] CG Peterfy et al. “Whole-organ magnetic resonance imaging score (WORMS) of the knee in osteoarthritis”. In: *Osteoarthritis and cartilage* 12.3 (2004), pp. 177–190.

[8] Frank W Roemer et al. “Semiquantitative assessment of focal cartilage damage at 3 T MRI: a comparative study of dual echo at steady state (DESS) and intermediate-weighted (IW) fat suppressed fast spin echo sequences”. In: *European journal of radiology* 80.2 (2011), e126–e131.

[9] Charles Peterfy and Manish Kothari. “Imaging osteoarthritis: magnetic resonance imaging versus x-ray”. In: *Current rheumatology reports* 8.1 (2006), pp. 16–21.

[10] PG Conaghan et al. “MRI and non-cartilaginous structures in knee osteoarthritis”. In: *Osteoarthritis and cartilage* 14 (2006), pp. 87–94.

[11] Kenji Suzuki. “Overview of deep learning in medical imaging”. In: *Radiological physics and technology* 10.3 (2017), pp. 257–273.

[12] Francesco Calivà et al. “Studying osteoarthritis with artificial intelligence applied to magnetic resonance imaging”. In: *Nature Reviews Rheumatology* 18.2 (2022), pp. 112–121.

[13] Yanming Guo et al. “Deep learning for visual understanding: A review”. In: *Neurocomputing* 187 (2016), pp. 27–48.

[14] Marc Kohli et al. “Implementing machine learning in radiology practice and research”. In: *American journal of roentgenology* 208.4 (2017), pp. 754–760.

[15] Ajay Shrestha and Ausif Mahmood. “Review of deep learning algorithms and architectures”. In: *IEEE access* 7 (2019), pp. 53040–53065.

[16] Aleksei Tiulpin et al. “Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach”. In: *Scientific reports* 8.1 (2018), pp. 1–10.

[17] Aleksei Tiulpin and Simo Saarakkala. “Automatic grading of individual knee osteoarthritis features in plain radiographs using deep convolutional neural networks”. In: *Diagnostics* 10.11 (2020), p. 932.

[18] Aleksei Tiulpin et al. “Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data”. In: *Scientific reports* 9.1 (2019), pp. 1–11.

[19] Dong Hyun Kim et al. "Can additional patient information improve the diagnostic performance of deep learning for the interpretation of knee osteoarthritis severity". In: *Journal of Clinical Medicine* 9.10 (2020), p. 3341.

[20] Berk Norman et al. "Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs". In: *Journal of digital imaging* 32.3 (2019), pp. 471–477.

[21] Kevin A Thomas et al. "Automated classification of radiographic knee osteoarthritis severity using deep neural networks". In: *Radiology. Artificial intelligence* 2.2 (2020).

[22] Bochen Guan et al. "Deep learning risk assessment models for predicting progression of radiographic medial joint space loss over a 48-MONTH follow-up period". In: *Osteoarthritis and cartilage* 28.4 (2020), pp. 428–437.

[23] Yassine Nasser et al. "Discriminative Regularized Auto-Encoder for early detection of knee osteoarthritis: data from the osteoarthritis initiative". In: *IEEE transactions on medical imaging* 39.9 (2020), pp. 2976–2984.

[24] Pingjun Chen et al. "Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss". In: *Computerized Medical Imaging and Graphics* 75 (2019), pp. 84–92.

[25] Valentina Pedoia et al. "3D convolutional neural networks for detection and severity staging of meniscus and PFJ cartilage morphological degenerative changes in osteoarthritis and anterior cruciate ligament subjects". In: *Journal of Magnetic Resonance Imaging* 49.2 (2019), pp. 400–410.

[26] Valentina Pedoia et al. "Diagnosing osteoarthritis from T2 maps using deep learning: an analysis of the entire Osteoarthritis Initiative baseline cohort". In: *Osteoarthritis and cartilage* 27.7 (2019), pp. 1002–1010.

[27] Aniket A Tolpadi et al. "Deep learning predicts total knee replacement from magnetic resonance images". In: *Scientific reports* 10.1 (2020), pp. 1–12.

[28] Jean-Baptiste Schiratti et al. "A deep learning method for predicting knee osteoarthritis radiographic progression from MRI". In: *Arthritis research & therapy* 23.1 (2021), pp. 1–10.

[29] Daichi Hayashi, Frank W Roemer, and Ali Guermazi. "Magnetic resonance imaging assessment of knee osteoarthritis: current and developing new concepts and techniques". In: *Clin Exp Rheumatol* 37.Suppl 120 (2019), pp. 88–95.

[30] Felix Eckstein et al. "Magnetic resonance imaging (MRI) of articular cartilage in knee osteoarthritis (OA): morphological assessment". In: *Osteoarthritis and cartilage* 14 (2006), pp. 46–75.

[31] Charles G Peterfy, Erika Schneider, and M Nevitt. "The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee". In: *Osteoarthritis and cartilage* 16.12 (2008), pp. 1433–1441.

[32] Getao Du et al. "Medical image segmentation based on u-net: A review". In: *Journal of Imaging Science and Technology* 64 (2020), pp. 1–12.

[33] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[34] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[35] Yifei Zhang. "A better autoencoder for image: Convolutional autoencoder". In: *ICONIP17-DCEC. Available online: http://users. cecs. anu. edu. au/Tom. Gedeon/conf/ABCs2018/paper/ABCs2018_paper_58. pdf (accessed on 23 March 2017)*. 2018.

[36] David J Hunter et al. "Evolution of semi-quantitative whole joint assessment of knee OA: MOAKS (MRI Osteoarthritis Knee Score)". In: *Osteoarthritis and cartilage* 19.8 (2011), pp. 990–1002.

[37] DJ Hunter et al. “Definition of osteoarthritis on MRI: results of a Delphi exercise”. In: *Osteoarthritis and cartilage* 19.8 (2011), pp. 963–969.

[38] Torsten Boegård et al. “Correlation between radiographically diagnosed osteophytes and magnetic resonance detected cartilage defects in the tibiofemoral joint”. In: *Annals of the rheumatic diseases* 57.7 (1998), pp. 401–407.

[39] MaryFran Sowers et al. “Associations of anatomical measures from MRI with radiographically defined knee osteoarthritis score, pain, and physical functioning”. In: *The Journal of Bone and Joint Surgery. American volume.* 93.3 (2011), p. 241.

[40] BJE de Lange-Brokaar et al. “Radiographic progression of knee osteoarthritis is associated with MRI abnormalities in both the patellofemoral and tibiofemoral joint”. In: *Osteoarthritis and Cartilage* 24.3 (2016), pp. 473–479.

[41] DJ Hart and TD Spector. “Kellgren & Lawrence grade 1 osteophytes in the knee—doubtful or definite?” In: *Osteoarthritis and cartilage* 11.2 (2003), pp. 149–150.

[42] Frank W Roemer et al. “An illustrative overview of semi-quantitative MRI scoring of knee osteoarthritis: lessons learned from longitudinal observational studies”. In: *Osteoarthritis and cartilage* 24.2 (2016), pp. 274–289.

[43] Daichi Hayashi et al. “Semiquantitative assessment of subchondral bone marrow edema-like lesions and subchondral cysts of the knee at 3T MRI: a comparison between intermediate-weighted fat-suppressed spin echo and Dual Echo Steady State sequences”. In: *BMC musculoskeletal disorders* 12.1 (2011), pp. 1–9.

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[45] Felix Ambellan et al. “Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the Osteoarthritis Initiative”. In: *Medical image analysis* 52 (2019), pp. 109–118.

[46] Stefan Klein et al. “Elastix: a toolbox for intensity-based medical image registration”. In: *IEEE transactions on medical imaging* 29.1 (2009), pp. 196–205.

[47] Nancy Johari and Natthan Singh. “Bone fracture detection using edge detection technique”. In: *Soft Computing: Theories and Applications*. Springer, 2018, pp. 11–19.

[48] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv:1312.6114* (2013).

[49] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems* 28 (2015).

[50] Joseph Redmon et al. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

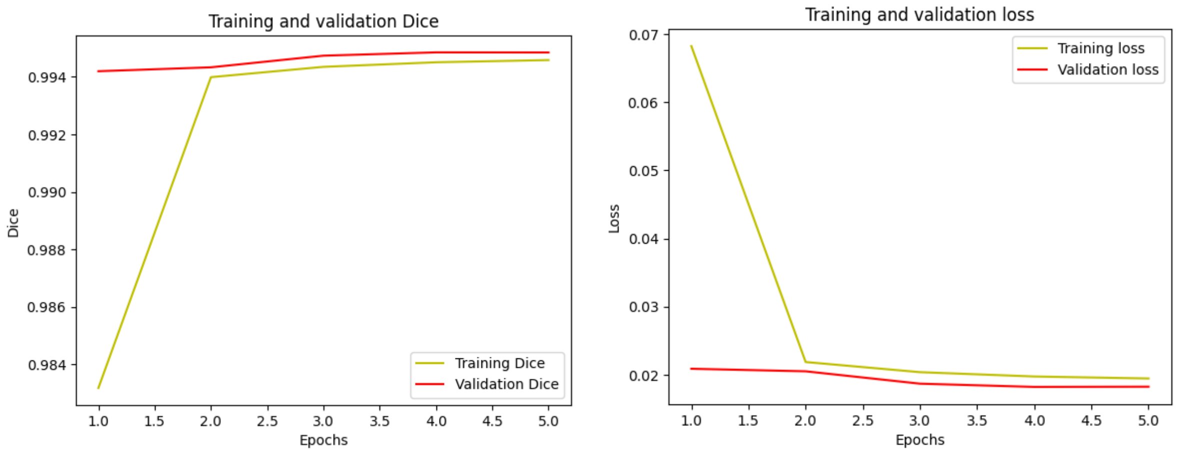
Appendix

U-Net

Batch size 4

	Batch Size 4			
Epoch	Training Dice	Validation Dice	Training Loss	Validation Loss
1	0.98318	0.99420	0.06826	0.02087
2	0.99399	0.99433	0.02185	0.02049
3	0.99435	0.99474	0.02036	0.01867
4	0.99451	0.99485	0.01973	0.01820
5	0.99458	0.99485	0.01944	0.01823

Table 1: Training and validation values for
Dice similarity coefficient and loss function for batch size 4



(a) Dice Similarity Coefficient

(b) Loss Value

Figure 1: Training and validation graphs for dice similarity coefficient and loss function for batch size 4

Batch Size 4		
Testing Dice	Testing Loss	Testing Accuracy
0.99509	0.01711	0.99334

Table 2: Testing Dice similarity coefficient ,loss and accuracy values for batch size 4

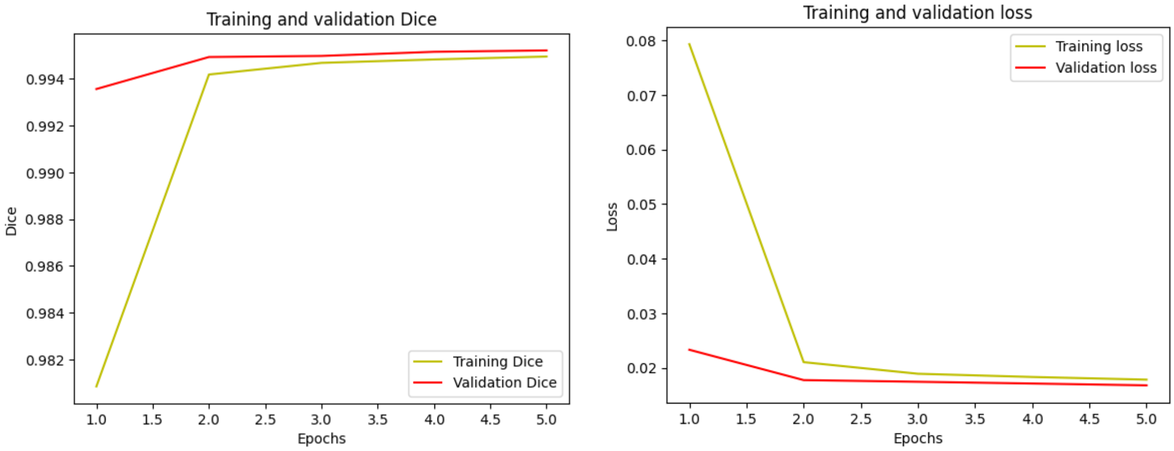
Intersection of Union - Batch Size 4					
Mean IoU	Background	Femoral Bone	Femoral Cartilage	Tibial Bone	Tibial Cartilage
0.88835	0.99292	0.97446	0.80423	0.96984	0.78605

Table 3: Testing intersection of union for batch size 4

Batch size 8

	Batch Size 8			
Epoch	Training Dice	Validation Dice	Training Loss	Validation Loss
1	0.98085	0.99355	0.07933	0.02328
2	0.99417	0.99492	0.02101	0.01772
3	0.99467	0.99497	0.01888	0.01742
4	0.99482	0.99515	0.01829	0.01709
5	0.99494	0.99521	0.01781	0.01676

Table 4: Training and validation values for
Dice similarity coefficient, loss and accuracy values for batch size 8



(a) Dice Coefficient

(b) Loss Value

Figure 2: Training and validation graphs for Dice similarity coefficient and loss function for batch size 8

Batch Size 8		
Testing Dice	Testing Loss	Testing Accuracy
0.99539	0.01607	0.99371

Table 5: Testing Dice similarity coefficient and loss function values for batch size 8

Intersection of Union - Batch Size 8					
Mean IoU	Background	Femoral Bone	Femoral Cartilage	Tibial Bone	Tibial Cartilage
0.89311	0.99196	0.96965	0.79797	0.96513	0.74119

Table 6: Testing Intersection of Union for Batch Size 8

Batch size 16

	Batch Size 16			
E	Training Dice	Validation Dice	Training Loss	Validation Loss
1	0.97066	0.99389	0.12606	0.02173
2	0.99424	0.99486	0.02067	0.01790
3	0.99480	0.99520	0.01831	0.01655
4	0.99493	0.99527	0.01784	0.01635
5	0.99516	0.99535	0.01684	0.01606

Table 7: Training and validation values for
Dice similarity coefficient and loss function values for batch size 16

Batch Size 16		
Testing Dice	Testing Loss	Testing Accuracy
0.99542168	0.0154375	0.99391770

Table 8: Testing Dice, loss and accuracy values for batch size 16

Batch Size 16 Intersection of Union					
Mean IoU	Background	Femoral Bone	Femoral Cartilage	Tibial Bone	Tibial Cartilage
0.9053756	0.9929245	0.9744608	0.8042309	0.9698405	0.7860468

Table 9: Testing intersection of union values for batch size 16

Resnet Results

Batch size: 4	Adam's learning rate: 0.001				Adam's learning rate: 0.01			
	ResNet-50		ResNet-50 + Clinical Data		ResNet-50		ResNet-50 + Clinical Data	
	AUC	**PR-AUC**	**AUC**	**PR-AUC**	**AUC**	**PR-AUC**	**AUC**	**PR-AUC**
orig	0.4603	0.2426	0.4837	0.2457	0.5462	0.2731	0.6335	0.3295
orig+reg	0.4277	0.2556	0.526	0.2835	0.6042	0.3194	0.6406	0.3613
orig+reg+bw	0.4036	0.2164	0.4551	0.2313	0.5645	0.3034	0.6302	0.345
orig+aug	0.5058	0.3161	0.6399	0.3632	0.5131	0.2787	0.6399	0.3643
orig+aug+reg	0.5775	0.3752	0.6458	0.3633	0.4837	0.3141	0.6374	0.3626

Table 10: ResNet-50 AUC & PR-AUC values for epochs number= 15,
batch size= 4 and Adam's optimizer learning rate= 0.001 (left) and 0.01 (right).
orig: trained on original data, orig+reg: trained on original data with kernel regularization
orig+reg+bw: trained in original data with kernel regularization and balanced weights
orig+aug: trained on augmented data, orig+aug+reg: trained on augmented data and kernel regularization

Batch size: 8	Adam's learning rate: 0.001				Adam's learning rate: 0.01			
	ResNet-50 output		ResNet-50 in combination with Clinical Data in Logistic Regression		ResNet-50 output		ResNet-50 in combination with Clinical Data in Logistic Regression	
	AUC	**PR-AUC**	**AUC**	**PR-AUC**	**AUC**	**PR-AUC**	**AUC**	**PR-AUC**
orig	0.5586	0.3829	0.5872	0.3678	0.5612	0.3014	0.6374	0.3387
orig+reg	0.4974	0.2617	0.5384	0.2912	0.5983	0.3115	0.6511	0.3786
orig+reg+bw	0.4128	0.2727	0.4401	0.2438	0.5534	0.2953	0.6198	0.3445
orig+aug	0.5241	0.2776	0.5729	0.2971	0.5559	0.3992	0.6426	0.3639
orig+aug+reg	0.5111	0.3054	0.5723	0.3157	0.5846	0.3912	0.6458	0.3669

Table 11: ResNet-50 AUC & PR-AUC values for epochs number= 15,
batch size= 8 and Adam's optimizer learning rate= 0.001 (left) and 0.01 (right).
orig: trained on original data, orig+reg: trained on original data with kernel regularization
orig+reg+bw: trained in original data with kernel regularization and balanced weights
orig+aug: trained on augmented data, orig+aug+reg: trained on augmented data and kernel regularization

DenseNet Results

Batch size: 4	Adam’s learning rate: 0.001				Adam’s learning rate: 0.01			
	DenseNet-121 output		DenseNet-121 in combination with Clinical Data in Logistic Regression		DenseNet-121 output		DenseNet-121 in combination with Clinical Data in Logistic Regression	
	AUC	**PR-AUC**	**AUC**	**PR-AUC**	**AUC**	**PR-AUC**	**AUC**	**PR-AUC**
orig	0.5481	0.2917	0.6204	0.3126	0.4603	0.2316	0.6269	0.3292
orig+reg	0.5273	0.2728	0.6426	0.3386	0.5618	0.2945	0.6556	0.3714
orig+reg+bw	0.5469	0.281	0.6074	0.3117	0.5521	0.2797	0.6198	0.3217
orig+aug	0.5033	0.2888	0.6426	0.3651	0.5241	0.3067	0.6406	0.364
orig+aug+reg	0.5358	0.3124	0.6393	0.3639	0.4746	0.2701	0.6419	0.3967

Table 12: DenseNet-121 AUC & PR-AUC values for epochs number= 15,
batch size= 4 and Adam’s optimizer learning rate= 0.001 (left) and 0.01 (right).
orig: trained on original data, orig+reg: trained on original data with kernel regularization
orig+reg+bw: trained in original data with kernel regularization and balanced weights
orig+aug: trained on augmented data, orig+aug+reg: trained on augmented data and kernel regularization

Batch size: 8	Adam’s learning rate: 0.001				Adam’s learning rate: 0.01			
	DenseNet-121 output		DenseNet-121 in combination with Clinical Data in Logistic Regression		DenseNet-121 output		DenseNet-121 in combination with Clinical Data in Logistic Regression	
	AUC	**PR-AUC**	**AUC**	**PR-AUC**	**AUC**	**PR-AUC**	**AUC**	**PR-AUC**
orig	0.4694	0.2461	0.5339	0.3134	0.6269	0.3711	0.6406	0.3653
orig_reg	0.4818	0.2789	0.6159	0.3747	0.5801	0.3103	0.6419	0.3522
orig_reg_bw	0.4915	0.2638	0.526	0.2669	0.5365	0.2818	0.6224	0.3242
aug	0.5358	0.3641	0.6543	0.3623	0.4941	0.2911	0.6348	0.3489
aug_reg	0.5501	0.3455	0.6497	0.3722	0.5026	0.2839	0.6393	0.3641

Table 13: DenseNet-121 AUC & PR-AUC values for epochs number= 15,
batch size= 8 and Adam’s optimizer learning rate= 0.001 (left) and 0.01 (right).
orig: trained on original data, orig+reg: trained on original data with kernel regularization
orig+reg+bw: trained in original data with kernel regularization and balanced weights
orig+aug: trained on augmented data, orig+aug+reg: trained on augmented data and kernel regularization

CVAE

Epochs = 100				Without Kernel Regularization			
				CVAE and Logistic Regression		CVAE +clinical and Logistic Regression	
BS	LR	Disc.Pen	Code.Dim	**AUC**	**PR-AUC**	**AUC**	**PR-AUC**
4	0.001	0.001	1000	0.625	0.3377	0.6378	0.3461
4	0.01	0.001	1000	0.6442	0.3625	0.6468	0.3654
8	0.001	0.001	1000	0.5968	0.3097	0.6128	0.3235
8	0.01	0.001	1000	0.5699	0.3519	0.5846	0.3597
4	0.001	0.01	1000	0.6468	0.3623	0.6564	0.3651
4	0.01	0.01	1000	0.6689	0.3463	0.6699	0.3589
8	0.001	0.01	1000	0.609	0.311	0.6359	0.3361
8	0.01	0.01	1000	0.55	0.2947	0.5699	0.3075
4	0.001	0.0005	1000	0.6205	0.318	0.6134	0.3256
4	0.01	0.0005	1000	0.6321	0.3396	0.6404	0.3395
8	0.001	0.0005	1000	0.6167	0.3227	0.6321	0.3404
8	0.01	0.0005	1000	0.5949	0.331	0.6096	0.3379

Table 14: CVAE AUC & PR-AUC values
for epochs number = 100 and latent space dimension = 100
without kernel regularization

Epochs = 100				With Kernel Regularization			
				CVAE and Logistic Regression		CVAE +clinical and Logistic Regression	
BS	LR	Disc.Pen	Code.Dim	**AUC**	**PR-AUC**	**AUC**	**PR-AUC**
4	0.001	0.001	1000	0.5705	0.2831	0.5994	0.3035
4	0.01	0.001	1000	0.5936	0.3067	0.6058	0.3252
8	0.001	0.001	1000	0.5872	0.3449	0.6141	0.3811
8	0.01	0.001	1000	0.5571	0.3148	0.566	0.3533
4	0.001	0.01	1000	0.525	0.2626	0.5571	0.2867
4	0.01	0.01	1000	0.5436	0.2785	0.5564	0.2938
8	0.001	0.01	1000	0.6064	0.4038	0.6218	0.4372
8	0.01	0.01	1000	0.5929	0.3297	0.6083	0.3511
4	0.001	0.0005	1000	0.5385	0.2706	0.5532	0.2868
4	0.01	0.0005	1000	0.6	0.3096	0.6135	0.3305
8	0.001	0.0005	1000	0.5904	0.4102	0.6026	0.4366
8	0.01	0.0005	1000	0.5814	0.3498	0.5962	0.368

Table 15: CVAE AUC & PR-AUC values
for epochs number = 100 and latent space dimension = 100
with kernel regularization