

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

**CLASSIFICATION OF KNEE OSTEOARTHRITIS SEVERITY
USING DEEP LEARNING WITH FULLY SUPERVISED
AND SEMI-SUPERVISED-BASED APPROACHES**

M.Sc. THESIS

İlknur AKTEMUR

Department of Computer Engineering

Computer Engineering Programme

JULY 2024

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

**CLASSIFICATION OF KNEE OSTEOARTHRITIS SEVERITY
USING DEEP LEARNING WITH FULLY SUPERVISED
AND SEMI-SUPERVISED-BASED APPROACHES**

M.Sc. THESIS

**İlknur AKTEMUR
(504201575)**

Department of Computer Engineering

Computer Engineering Programme

Thesis Advisor: Assoc. Prof. Dr. Ilkay OKSUZ

JULY 2024

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

**DERİN ÖĞRENME İLE TAMAMEN DENETİMLİ
VE YARI-DENETİMLİ YAKLAŞIM TABANLI
DİZ OSTEOARTRİT ŞİDDETİNİN SINIFLANDIRILMASI**

YÜKSEK LİSANS TEZİ

**İlknur AKTEMUR
(504201575)**

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Assoc. Prof. Dr. Ilkay OKSUZ

TEMMUZ 2024

İlknur AKTEMUR, a M.Sc. student of ITU Graduate School student ID 504201575 successfully defended the thesis entitled “CLASSIFICATION OF KNEE OSTEOARTHRITIS SEVERITY USING DEEP LEARNING WITH FULLY SUPERVISED AND SEMI-SUPERVISED-BASED APPROACHES”, which he/she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor : **Assoc. Prof. Dr. Ilkay OKSUZ**
Istanbul Technical University

Jury Members : **Prof. Dr. Mustafa Alper SELVER**
Dokuz Eylül University

Asst. Prof. Dr. Yusuf Hüseyin SAHİN
Istanbul Technical University

Date of Submission : **24 May 2024**
Date of Defense : **02 July 2024**





To my family and friends,



FOREWORD

I would like to show my appreciation and gratitude to my supervisor, Dr. Ilkay OKSUZ. I am grateful for his deep knowledge, detailed guidance and support. I would like to thank Caner OZER for all his support and experience throughout my thesis research process. I would also like to thank the PIMI Lab members for sharing valuable information. Finally, but surely not least, I would like to express my sincere thanks to my family and friends who have always been there for me. Their constant and endless support led me to conduct this study.

July 2024

İlknur AKTEMUR
(Machine Learning Engineer)

TABLE OF CONTENTS

	<u>Page</u>
FOREWORD	ix
TABLE OF CONTENTS	xii
ABBREVIATIONS	xiii
LIST OF TABLES	xv
LIST OF FIGURES	xviii
SUMMARY	xix
ÖZET	xxi
1. INTRODUCTION	1
1.1 Purpose of the Thesis	1
1.2 Outline	3
2. BACKGROUND	5
2.1 Medical Background	5
2.1.1 Definition of the knee osteoarthritis and importance	5
2.1.2 Stages of the disease	7
2.1.3 Imaging examinations for osteoarthritis	8
2.2 Machine Learning Background	11
2.2.1 Supervised learning	11
2.2.1.1 Classification	11
2.2.1.2 Regression	12
2.2.2 Transfer learning and self-supervised learning for data scarcity	12
2.2.2.1 Contrastive learning	13
2.2.2.2 Self-supervised contrastive learning	15
2.2.2.3 Supervised contrastive learning	15
2.2.3 Computer vision	16
2.2.3.1 Convolutional neural networks (CNN)	16
2.2.3.2 Vision transformers	17
3. LITERATURE REVIEW	19
3.1 Computer-Aided Diagnosis in Disease Detection	19
3.2 Machine Learning Techniques for Knee Osteoarthritis	20
3.3 Deep Learning Approaches for Predicting Knee Osteoarthritis	21
3.4 Semi-Supervised Learning and Contrastive Learning Based Methods	25
4. SUPERVISED LEARNING METHODS IN KNEE OA SEVERITY CLASSIFICATION	29
4.1 Model Architectures	29
4.2 Loss Functions and Evaluation Metrics	33
4.2.1 Loss function	34
4.2.1.1 Cross entropy loss function	34
4.2.1.2 Ordinal loss function	34
4.2.2 Evaluation metrics	36
4.2.2.1 Weighted accuracy	36
4.2.2.2 F1 score	36
4.2.2.3 MAE	38
4.2.2.4 sa-RPS	38
4.3 Experimental Setups	39
4.3.1 Parameter setup	40
4.3.2 Datasets	41
4.3.3 Detail implementation	42
4.3.4 Data preprocessing	42
4.3.4.1 One-stage fine-tuning	43
4.3.4.2 Two-stage fine-tuning	43
4.4 Results and Analysis for Supervised Learning Methods	44
4.4.1 Computer vision model architecture comparison	47

4.4.2	Loss functions comparison	49
4.4.3	Paradigm for fine-tuning comparison	50
5.	SELF- SUPERVISED LEARNING METHODS IN KNEE OA SEVERITY CLASSIFICATION	53
5.1	Proposed Semi-Supervised Paradigm	53
5.2	Experimental Setups	56
5.2.1	Parameter setup	56
5.2.2	Datasets	57
5.2.3	Data preprocessing	57
5.2.4	Detail implementation	58
5.3	Results and Analysis for Semi- supervised Learning Methods	59
6.	CONCLUSIONS	61
6.1	Answers	61
6.2	Summary and Contributions	62
6.3	Limitations and Future Work	64
	REFERENCES	65
	CURRICULUM VITAE	74



ABBREVIATIONS

AP	: Anteroposterior
CAD	: Computer-Aided Diagnostic
CE	: Cross Entropy
CL	: Contrastive Learning
CNN	: Convolutional Neural Network
CT	: Computed Tomography
GAP	: Global Average Pooling
Grad-CAM	: Gradient Weighted Class Activation Mapping
GRN	: Global Response Normalization
ICA	: Independent Component Analysis
JSN	: Joint Space Narrowing
KL	: Kellgren-Lawrence
MRI	: Magnetic Resonance Imaging
MAE	: Mean Absolute Error
MLP	: Multi Layer Perceptron
MLR	: Multivariate Linear Regression
MOST	: Multicenter Osteoarthritis
OA	: Osteoarthritis
OAI	: Open Archives Initiative
PA	: Posteroanterior
RAdam	: Rectified Adam
ReLU	: Rectified Linear Unit
ROC	: Receiver Operating Characteristic
ROI	: Region of Interest
SOTA	: State of the Art
SSL	: Semi-Supervised Learning
SwAV	: Swapping Assignments Between Views
TF	: Tibial Femoral
TKR	: Total Knee Replacement
USG	: Ultrasound
ViT	: Vision Transformer



LIST OF TABLES

	<u>Page</u>
Table 2.1 : Key Properties of Imaging Techniques for Knee OA Diagnosis [1]...	10
Table 4.1 : Data distributions among classes and subsets for the public OAI dataset and the in-house hospital dataset.	42
Table 4.2 : The performances of the models that were fine-tuned once using the OAI dataset.	46
Table 4.3 : The performances of the models that were fine-tuned once using the in-house hospital dataset.	46
Table 4.4 : The performances of the models that were two-stage fine-tuned using the in-house hospital dataset.	46
Table 4.5 : Comparison of two-stage fine-tuned and one-stage fine-tuned ConvNeXt-Base model using in-house hospital dataset.	47
Table 5.1 : Data distribution between classes in progressively reduced training data for supervised fine-tuning stage.	58
Table 5.2 : Performance comparisons between our proposed SimCLR-based model and the baseline model on the OAI test dataset.	59



LIST OF FIGURES

	<u>Page</u>
Figure 2.1 : Schematic drawing of an osteoarthritic joint [2].	6
Figure 2.2 : Sample X-ray images are arranged from 0 to 4 based on KL stages [3].	8
Figure 2.3 : Contrastive learning paradigm that bringing similar examples closer, pushing dissimilar ones further apart in representation space.	14
Figure 2.4 : Supervised vs. self-supervised contrastive learning [4].	15
Figure 2.5 : An example of CNN architecture [5].	17
Figure 2.6 : Vision transformer(ViT) overview [6].	18
Figure 3.1 : Proposed classification pipeline in [7] study.	22
Figure 3.2 : Flowchart of proposed key-patch exchange strategy in [8].	24
Figure 4.1 : Residual block [9]. Skip-connections or residual-connections that is a solution for vanishing gradient in deep networks.	30
Figure 4.2 : Effect of width, depth and resolution on model efficiency. The points on the lines represent different models with different width coefficient (w) [10].	32
Figure 4.3 : Block designs for a ResNet, a Swin Transformer, a ConvNeXt V1 and a ConvNeXt V2 [11,12].	33
Figure 4.4 : Ordinal loss calculation schema.	35
Figure 4.5 : Calculation process of sa-RPS score.	39
Figure 4.6 : In the two-stage fine-tuning schema, fine-tuning stages are seen for 2 different datasets. In stage 1(a), a pre-trained model with ImageNet weights is fine-tuned with larger data from OAI dataset, and the weights of this model are transferred to the model in stage 2(b) and the all layers are frozen except the last 100 layers. In stage 2, it was fine-tuned again with less data from our in-house dataset, using the same pre-processing steps and the same model architecture as in the first stage.	45
Figure 4.7 : Attention map comparison using Grad-CAM technique.	48
Figure 4.8 : Localizations produced using the Grad-CAM method on the ConvNeXt-Base model. This model was obtained by a two-stage fine-tuning strategy and ordinal loss on hospital dataset. The first and third row are samples of knee X-rays and the second and fourth row are their corresponding overlaid classification activation maps. Joint space narrowing and osteophyte formation were successfully captured when the model predicted the grade of knee OA.	51
Figure 5.1 : Overview of SimCLR [13].	54
Figure 5.2 : Proposed semi-supervised framework based on contrastive learning.	55

Figure 5.3 : Accuracy and F1-score for our proposed SimCLR based model and baseline model on test set. **60**



CLASSIFICATION OF KNEE OSTEOARTHRITIS SEVERITY USING DEEP LEARNING WITH FULLY SUPERVISED AND SEMI-SUPERVISED-BASED APPROACHES

SUMMARY

Osteoarthritis (OA) occurs when the cartilage tissue, a smooth substance found at the ends of bones, thins over time and even disappears. Symptoms such as swelling, pain, and restricted movement in the affected knees significantly diminish the quality of life for individuals and pose a significant health problem that limits their daily activities. To alleviate the damages caused by osteoarthritis and slow down the progression of the disease, accurate determination of the disease's stages is essential. Cartilage degeneration, wear, reduced joint space, and bone changes seen on X-rays offer indirect insights into knee OA stages. This situation causes disagreement among radiologists and hinders early detection. Recently, there has been an increase in studies focusing on accelerating early diagnosis and supporting radiologists' decisions by automatically determining disease stages with deep learning-based methods. However, high-performance models are still needed. The success of deep learning often depends on large-scale labeled data. However, the data labeling phase of medical images is costly and laborious. Within the scope of this thesis, we recognized this challenge, we proposed fully supervised and semi-supervised methods suited for scenarios with abundant or limited labeled data.

Firstly, we fine-tuned models like ResNet, DenseNet, EfficientNet, ResNeXt, ConvNeXt trained on the ImageNet using the publicly available OAI dataset in fully supervised learning part. Also during fine-tuning, we used the categorical cross-entropy loss function and the ordinal loss function. We called this process one-stage fine-tuning. Fine-tuning with the ordinal loss function yielded the best results on the OAI test data, with a 73.91% accuracy score and a 73.77% F1-score achieved by the ConvNeXt V2-Tiny model. The ConvNeXt V1-Base model provided the best result in Mean Absolute Error (MAE) metric with a score of 0.28.

For scenario with a limited amount of labeled data gathered from a single center, we employ a two-stage fine-tuning approach within the realm of fully supervised methods. We re-fine-tune models that obtained from first stage-fine tuning using smaller-scale hospital data this time. The two-stage fine-tuning method has yielded higher performance scores across all models. The best model performance belongs to the ConvNeXt V1-Base model again, with a 70.86% accuracy score, 71.35% F1-score, and a 0.30 MAE score.

For scenarios where unlabeled data more than labeled data, we propose a semi-supervised approach utilizing both types of data. Here, we integrate representations learned through a pre-training process based on Contrastive Learning (CL) using unlabeled X-ray data into a supervised fine-tuning process conducted with a

small amount of labeled data. We also examine model robustness by employing semi-supervised learning, gradually reducing labeled data by 25% in each experiment. Our model trained with the semi-supervised learning approach outperforms the baseline model with a 64.16% accuracy score, 64.72% F1 score, and 0.42 MAE score. Furthermore, as the labeled data decreases, the gap between them widens for all performance metrics.



DERİN ÖĞRENME İLE TAMAMEN DENETİMLİ VE YARI-DENETİMLİ YAKLAŞIM TABANLI DİZ OSTEOARTRİT ŞİDDETİNİN SINIFLANDIRILMASI

ÖZET

Osteoartrit (OA) dünyada en sık görülen kemik hastalıklarındandır ve diz osteoartriti osteoartritin en sık rastlanan formudur. Kemik uçlarında kaygan kıvamı ile eklemden sürtünmeyi azaltan kıkırdak dokusunun zamanla incilmesi ve hatta yok olması kemiklerin doğrudan birbirine sürtünmesine sebep olur. Bunun sonucunda eklemlerde meydana gelen şişlik, ağrı ve hareket kısıtlılığı gibi belirtiler kişilerin yaşam kalitesini önemli ölçüde bozan ve günlük aktivitelerini engelleyen bir sağlık sorunu oluşturur. Bu durum; hastaların tıbbi tedavi maliyetleri, kişilerin günlük aktivitelerini yerine getirememesiyle ortaya çıkan bakım ihtiyacı, iş gücü kaybı, engelli yardımları ve ekonomik verimlilik kaybı gibi farklı şekillerde toplum için büyük bir ekonomik yükü de beraberinde getirmektedir. Osteoartritin kişilerde ve toplumda sebep olduğu bu zararları hafifletmek için hastalığın ilerlemesini yavaşlatmak veya en azından gelecekteki sakatlık etkisini azaltmak mümkündür.

Bu, hastalığın aşamalarının sistematik olarak teşhis edilmesini gerektirmektedir. Diz OA'sının teşhisi için Radyografi (X-Işını), Manyetik Rezonans Görüntüleme (MRI), Ultrason (USG) ve Bilgisayarlı Tomografi (BT) gibi tıbbi görüntüleme yöntemleri kullanılabilir. Diz eklemlerinin MRI cihazlarıyla elde edilen üç boyutlu görüntüsü, dizdeki yapısal bozuklukları detaylı olarak yansıtılabilir. Ancak, MRI cihazlarının yalnızca büyük tıp merkezlerinde bulunması ve yüksek maliyeti sebebiyle diz OA'sı teşhisinde kullanımı çok yaygın değildir. Bu sebeple diz OA teşhisinde genellikle radyografi tercih edilmektedir.

Fakat OA değişikliklerini saptamaya çalışırken radyografinin yetersiz olduğu durumlar vardır. Bunun nedenleri arasında, radyografi ile gözlemlenebilen kıkırdak dejenerasyonu, kıkırdak aşınması, eklem aralığının daralması ve kemik değişikliklerinin değerlendirilmesinin, diz OA evresi hakkında dolaylı bilgi sunması yer alır. Bu nedenle, radyografinin doğru bir şekilde yorumlanabilmesi için deneyimli bir radyolog gereklidir. Ayrıca, eklemdaki değişikliklerin değerlendirilmesi üç boyutlu bir problem olduğu halde, radyografi yalnızca iki boyutlu görüntüler sağlar.

Radyografinin yetersiz olduğu durumlar dışında, diz OA'sı tespitini zorlaştıran başka nedenler de vardır. Örneğin, OA'nın evrelerini belirlemek için net bir derecelendirme sistemi bulunmaması, tanı sürecinin radyologların kişisel değerlendirmelerine bağlı kalmasına neden olmaktadır. Yaygın olarak kullanılan Kellgren-Lawrence (KL) ölçeği, yarı nicel bir sistemdir ve bu da radyologlar arasında farklı yorumlara yol açar. Bu tutarsızlıklar, hastalığın evresinin doğru bir şekilde belirlenmesini zorlaştırır ve tedavi sürecini etkiler. Sonuç olarak, dünya genelinde milyonlarca insan bu sorundan etkilenmektedir. Bu nedenle, diz OA'nın evresini belirlemek ve erken tanı koymak klinik pratikte zordur.

Öte yandan şu anda hastalığın yaygınlığı, hastalığın doğru bir şekilde tespit edilmesi ve boyutunun ölçülebilmesi konusunda acil bir ihtiyaç doğurmaktadır. Son zamanlarda, yapay zeka destekli bilgisayar uygulamalarının gelişmesiyle literatürde bu ihtiyaca yönelik çalışmalar yapılmaktadır. Bu çalışmalarla hastalığın evresi ile ilgili doktorların yaşadıkları fikir ayrılıklarını azaltılması, hastalığın evrelerinin doğru tanımlanarak erken teşhisinin hızlandırılması amaçlanmıştır. Bu çalışmada, derin öğrenme tabanlı yöntemlerin radyalogları kararlarında desteklemek için objektif bir araç olarak kullanılabileceğine ve diz OA tanısının önemli ölçüde iyileştirilebileceğine inanılmıştır. Bu sebeple tez kapsamında, diz X-ray görüntülerinden diz OA'sının evrelerini belirlemek üzere, derin öğrenme tabanlı yöntemler sunulmaktadır.

Görüntü tanıma ve sınıflandırma problemlerinde yaygın olarak kullanılan Evrişimli Sinir Ağları (CNN) çeşitli hastalık teşhisi sistemlerinde çığır açıcı sonuçlar ortaya koymuştur. Bu modellerin pek çok alanda insan düzeyine erişen performansları, klinik uygulamalarda kullanımının yakın gelecekte daha da yaygınlaşacağını açıkça göstermektedir. CNN'ler, teşhise ait özellikleri görüntülerden otomatik olarak öğrenir. Fakat bu modellerin eğitilmeleri esnasında görüntünün doğru özelliklerine odaklanması çoğu zaman büyük etiketli veri kümelerinin varlığına bağlıdır. Bu durum, başarılı otomatik teşhis sistemleri oluşturulmasında en büyük engellerden biridir. Çünkü tıbbi görüntülerin etiketlenmesi sadece uzmanlar tarafından gerçekleştirilebilen maliyetli bir işlemdir. Bu çalışmada, derin öğrenme yöntemlerini farklı veri senaryolarına uyarlayanın, literatürdeki çeşitli veri senaryoları için daha başarılı modeller geliştirilmesine katkı sağlayacağı düşünülmüştür. Bu sebeple bu çalışmada sunulan yöntemler etiketli veri sayısının yeterli olduğu ve yeterli olmadığı senaryolar üzerine inşa edilmiştir.

İlk olarak, kullanılan etiketli veri sayısının nispeten yeterli olabileceği durumlar için, tamamen denetimli öğrenme yöntemleri uygulanmıştır. Bu kısımda, literatürde eskiden beri yaygın olarak kullanılan ResNet, DenseNet , EfficientNet gibi modellerin yanında yanında ConvNeXt gibi daha güncel modellerin de ImageNet verisi ile eğitilmiş versiyonlarını halka açık OAI veriseti ile ince ayar(fine-tune) yapılmıştır. Bunun yanında ince ayar esnasında, kategorik çapraz entropi kaybı fonksiyonu ile sıralı kayıp fonksiyonu kullanılmıştır. Sıralı kayıp fonksiyonu ile ince ayar yapılmış ConvNeXt V2-Tiny modeli , OAI test veri üzerinde %73.91 doğruluk skoru ve %73.77 F1-skoruyla en iyi sonuçları vermiştir. ConvNeXt V1-Base modeli 0.28 skoruyla Ortalama Mutlak Hatası (MAE) metriğinde en iyi sonucu vermiştir.

Tek bir merkezden toplanmış, kısıtlı miktarda etiketli verinin olduğu senaryo için ise yine tamamen denetimli yöntemler kapsamında iki aşamalı ince-ayar yöntemi kullanılmıştır. Çoklu merkezden toplanan, daha çeşitli X-ray verilerini bulunduran ve daha büyük boyuttaki halka açık OAI veriseti, KL evreleri hakkında daha fazla temsil sunmaktadır. Bu nedenle, mevcut OAI veriseti kullanılarak ince ayar yapılmış ve önemli temsiller içeren modellere, daha küçük boyuttaki hastane verilerini kullanarak tekrar ince ayar yapılmıştır. OAI verisetinden öğrenilen temsillerin kaybolmaması için, ikinci aşamada modelin son yüz katmanı hariç tüm katmanları dondurulmuştur. Son yüz katmanda da yeni veriye ait özelliklerin öğrenilmesi sağlanmıştır. İki-aşamalı ince-ayar metodu, tüm modeller için daha yüksek performans skorları sağlamıştır. En iyi model performansı, %70.86 doğruluk skoru, %71.35 F1-skoru ve 0.30 MAE skoru

ile yine ConvNeXt V1-Base modeline aittir. Bu adımda, modellerin performansını değerlendirmek için ek olarak, sınıflar arası olasılık dağılımını da dikkate alarak daha şeffaf bir performans ölçüsü olarak kabul edilen güncel sa-RPS metriği de kullanılmıştır. ConvNeXt V1-Base modeli, sa-RPS metriği için de 0.029 skoruyla, diğer modellere kıyasla en iyi performansı sergilemiştir.

Etiketsiz verinin etiketli veriden çok olduğu senaryolar için ise, her iki tür veriyi kullanarak yarı denetimli bir yöntem önerilmiştir. Bu aşamada, etiketsiz X-ray verilerini kullanarak Karşılaştırmalı Öğrenme'ye(CL) dayanan bir ön eğitim sürecinde öğrenilen temsilleri, küçük miktarda etiketli veri ile gerçekleştirilen denetimli ince ayar sürecine entegre edilmiştir. CL ile ön eğitilmiş modellerin performansı ImageNet ağırlıklarını barındıran ConvNeXt V1-Base temel modeli ile kıyaslanmıştır. Ayrıca modelin küçük, etiketli veri kümeleriyle eğitildiğinde sağlamlığını test etmek için etiketli veriyi kademeli olarak azaltarak (her deneyde %25) yarı denetimli öğrenme yaklaşımıyla oluşturulan model performansı biraz daha araştırılmıştır. Yarı denetimli öğrenme yaklaşımıyla eğitilen model, %64.16 doğruluk skoru, %64.72 F1-skoru ve 0.42 MAE skoru ile temel modele karşı üstünlük sağlamıştır. Ayrıca etiketli veri sayısı azaldıkça, tüm performans metriklerinde aralarındaki fark daha da açılmıştır.



1. INTRODUCTION

Knee Osteoarthritis (OA) affects over 80% of individuals over the age of 55 and stands as one of the most prevalent bone disorders. Over the past two decades, the incidence of knee OA in the United States alone has doubled in men and tripled in women. Consequently impacting approximately 250 million individuals globally. This escalating prevalence underscores a significant societal burden. OA manifests as the gradual degradation and breakdown of the cartilage that protects the bones within the joints. In the early stages of the knee OA, symptoms such as pain and tenderness typically seen after intense activities, as the deterioration of the cartilage surface is minimal. However, in more severe stages of Knee OA, there is notable cartilage loss, deformities in shape, inflammation, fluid accumulation, and the formation of bone growths (osteophytes). So, even during periods of rest, pain and restricted movement in the knee are observed. Treatment approaches for steoarthritis vary across its different stages. In the early stages, methods such as exercises, analgesics, and physical therapy may be employed to alleviate pain and prevent its progression. However, in more severe stages, surgical intervention known as joint replacement, wherein an artificial prosthesis replaces the damaged joint, is applied. The diagnosis of the disease is typically made by clinicians based on the individual's symptoms and radiographic findings.

1.1 Purpose of the Thesis

The subjective nature of radiographic findings frequently results in disparities in evaluations by radiologists, thereby impacting the precision of determining knee OA grade and treatment decisions. Therefore, the automated detection of knee OA using knee X-ray images has become a research focus in recent years. These investigations have sought to mitigate radiologists' discrepancies regarding disease staging and expedite disease diagnosis. However, accurately predicting the stage of knee OA

using knee X-ray data has posed a challenging problem for researchers due to the minimal differences in adjacent stages, and the model performances in the literature is still not satisfactory. One of the most influential factors affecting the performance of these models is the utilized data. Therefore, considering and implementing different methods for various data scenarios is crucial for enhancing the performance of diagnostic models. Based on this premise, in this study, we evaluate supervised and semi-supervised deep learning methods to obtain a robust approach for automatic detection of knee OA stage from knee X-ray images. For situations where a sufficient amount of labeled data is available, we aim to construct well-performing models by exploring the impact of model architecture and loss function on model performance, using different supervised deep learning models and loss functions. When labeled knee X-ray data is collected from a single center, we investigate the effect of utilizing another labeled knee X-ray datasets collected from multiple centers in the literature on building a robust model.

Generating large-scale labeled data is a specialized, time-consuming, and costly process. On the flip side, the scarcity of labeled data constrains the model's capacity to generalize while also increases the risk of overfitting. However, unlabeled medical data is abundant, easy to access, and cheap. In this study, we proposed a semi-supervised pipeline based on contrastive learning to enhance the performance of knee OA stage classification when labeled data is insufficient and unlabeled data is available. Our objective is to find answer for the following research inquiries at the end of this report:

- **Question 1** How do different deep learning architectures perform in fully supervised learning?
- **Question 2** How much a deep model is sensitive to choosing the right loss function for the severity of knee OA classification task?
- **Question 3** What is the effect of the two-stage fine-tuning in KL based knee OA severity classification when we have small size labeled data?
- **Question 4** How can we leverage unlabeled images to train a deep model to have a robust model?

1.2 Outline

We organize the thesis as follows:

- **Chapter 2** provides background information about clinical importance and machine learning methods.
- **Chapter 3** explores existing research on computer-aided diagnosis systems, focusing on studies pertaining to knee osteoarthritis and other disease diagnostic tasks within the realm of artificial intelligence
- **Chapter 4** presents proposed fully supervised learning methods in knee OA severity classification to boost model performance.
- **Chapter 5** demonstrates proposed semi-supervised classification pipeline based on contrastive learning to overcome limitations caused by the scarcity of labeled data.
- **Chapter 6** summarizes the thesis contributions, addresses the research questions, and explores avenues for future work.

2. BACKGROUND

In this chapter, we present the medical background for the clinical aspect of our research and the machine learning background for our methods in the machine learning.

2.1 Medical Background

In this section, there are three subsections. Firstly, in Chapter 2.1.1, we discussed the pathogenesis of knee OA and the symptoms involved. Secondly, in Chapter 2.1.2, we explored the staging of knee OA disease according to the KL grade. Lastly, in Chapter 2.1.3, we described the imaging modalities used for the detection of knee OA.

2.1.1 Definition of the knee osteoarthritis and importance

Cartilage, a tissue present in both human and animal organisms, serves a pivotal role in the physiological framework [14]. Possessing a composition characterized by flexibility, durability, and a slippery consistency, it acts as a connective medium between bone extremities, mitigating friction within joints, furnishing structural support, and safeguarding certain organ configurations. Comprised of collagen fibers interwoven with proteoglycan molecules, cartilage manifests its pivotal attributes, facilitating and safeguarding the mobility of joints [15]. Its significance in human health lies in the preservation of joint integrity and the provision of structural reinforcement, enabling bone articulation devoid of abrasive contact. However, in the case of arthritis, the cartilage becomes thinner, cracks form, and the cartilage tissue may completely disappear over time [16]. This situation leads to bones rubbing directly against each other, resulting in symptoms such as pain, swelling, and restricted joint movement. Knee OA represents the most common type of arthritis, constituting a significant health concern that markedly impairs the quality of life and restricts daily activities [17]. Knee OA is characterized by pain and stiffness in the joint during movement or after extended periods of inactivity [18]. Inflammation and swelling

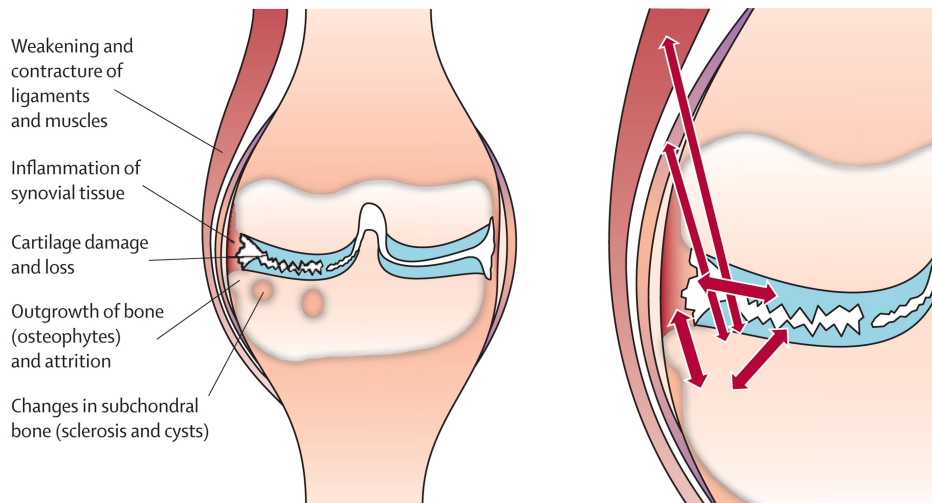


Figure 2.1: Schematic drawing of an osteoarthritic joint [2].

of the tissues surrounding the joint may occur. Biomarkers caused by knee OA in the joint are shown in Figure 2.1. Furthermore, knee OA leads to restricted joint mobility, notably pronounced in the morning or after prolonged immobilization. In advanced stages of knee OA, auditory sensations such as crunching, clicking, or crackling may accompany joint movement. These typically stem from cartilage tissue degradation or friction between joint surfaces, serving as symptomatic indicators of joint changes. Pain and stiffness in the joints may render even simple daily tasks challenging, consequently diminishing functionality. Moreover, at advanced stages of knee osteoarthritis, joint deformities contribute to physical disabilities.

If diagnosed early and appropriate treatment is administered, the progression of osteoarthritis can be slowed, and symptoms can be managed [19]. The incidence of knee OA has increased recently. In the last 20 years, knee OA prevalence has doubled in men and tripled in women in the United States, affecting around 250 million people globally [20]. The incidence of knee OA increases with aging. It has been stated that it affects 80% of individuals over the age of 55 [21]. Apart from age, factors such as excess weight, congenital curvature of bones and joints, repetitive physical movements that put pressure on the knee, knee injuries, and genetics also play a crucial role in the development of knee osteoarthritis.

2.1.2 Stages of the disease

Diagnosing the stages of knee OA is a critical step in comprehending the disease's progression, devising suitable treatment regimens, and monitoring its advancement [22]. During the early stages of knee OA, symptoms are milder, and joint damage is minimal. Conservative treatment alternatives like medication, exercise programs, physical therapy, and weight management are typically employed during this period for pain alleviation [23].

However, as OA advances, joint damage becomes more severe, accompanied by heightened symptomatology. In such advanced stages, more aggressive treatment modalities, including surgical interventions like prosthesis surgery, are warranted. Hence, accurate staging of the disease is imperative for formulating appropriate treatment strategies. The stages of knee OA are typically determined based on factors including clinical evaluation and radiological imaging tests [1]. During clinical evaluation, symptoms such as the severity of joint pain, limitation of movement, swelling, tenderness, and joint deformities are assessed. Imaging techniques can also reveal signs of osteoarthritis, such as reduced joint space, bone spurs, bone hardening, and intra-articular fluid accumulation [24].

To provide a clear definition of the disease stages for early diagnosis and treatment planning, various methods are employed, with one of the prominent ones being the KL system [25]. This grading method was developed to categorize the extent of joint damage through radiographic images and marked the initial formal endeavor to establish a radiographic classification system for OA. It was initially proposed in 1957 by Dr. John Kellgren and Dr. Jeffrey Lawrence, who delineated five distinct stages for classification. In assessing intraobserver and interobserver reliability, they examined evaluations based on plain radiographs. The high correlation coefficients they achieved rendered the KL classification the most extensively employed clinical tool for radiographic diagnosis of OA [24].

Grade 0 NO OA	Grade 1 Doubtful OA	Grade 2 Mild OA	Grade 3 Moderate OA	Grade 4 Severe OA
				
No Osteophytes	Possible Osteophytes	Definite Osteophytes	Moderate Osteophytes	Large Osteophytes
No JSN	Doubtful JSN	Possible JSN	Definite JSN	Great JSN

Figure 2.2: Sample X-ray images are arranged from 0 to 4 based on KL stages [3].

The radiographs were graded on a scale from 0 to 4, reflecting the increasing severity of OA. Grade 0 represents the absence of OA, whereas Grade 4 denotes severe OA [26]. X-ray images depicting each stage, along with their respective characteristics, are illustrated in Figure 2.2

2.1.3 Imaging examinations for osteoarthritis

Medical imaging techniques such as radiographs (X-ray), magnetic resonance imaging (MRI), ultrasound (USG), and computed tomography (CT) can be employed for knee OA detection; however, X-ray is generally favored for knee OA diagnosis [1]. This preference stems from its lower cost compared to other imaging modalities. Radiography can also reveal space between joint bones, cartilage erosion, and osteophytes (bone spurs) at the edges of the joint to indicate the presence of OA.

In radiography, bone calcium absorbs X-ray radiation, rendering bones appearing white. The patient may be positioned in various configurations during X-ray imaging, including supine, sitting, fully extended, semi-flexed, and weight-bearing positions. In the osteoarthritic joint, bone tissue may become dense, hardened and thickened due to wear of the cartilage. In this case, the bones appear more prominent and harder on X-rays. The anteroposterior (AP) and lateral views, obtained from various positions including supine, sitting, and standing, provide a comprehensive assessment of both patellofemoral joints in commercial imaging. These positions include fully extended, partially bent, without bearing weight, and with weight-bearing, aiding in identifying any cartilage deterioration in the inner or outer parts of the knee joint. Moreover,

the posteroanterior weight-bearing Rosenberg view, with the knee positioned at a 45° angle, offers increased sensitivity in detecting joint space narrowing (JSN).

Radiography serves as the gold standard for diagnosing knee OA, whereas MRI, CT, and ultrasound are not utilized as standard practices in clinical settings but rather considered advanced imaging techniques [1]. These advanced modalities boast heightened sensitivity in detecting early-stage disease progression, often averting substantial degeneration. MRI operates on the principles of magnetic wave theory. During an MRI scan, the patient lies on their back inside the MRI machine. This imaging modality offers heightened sensitivity, capable of identifying reactive bone edema, soft tissue inflammation, degenerated cartilage, and other soft tissue damage associated with OA, through the acquisition of high-contrast longitudinal and cross-sectional images.

CT is a prevalent imaging modality utilized for detailed examination and detection of internal body structures [27]. During CT imaging, the body is scanned with X-rays in thin sections, which are subsequently reconstructed into a three-dimensional image by computer software. CT scans play a crucial role in diagnosing knee osteoarthritis by accurately revealing bone structures and osteophytes. These scans can be performed with the subject bearing weight or without bearing any weight at all.

Ultrasound scanning, which employs ultrasound waves to examine soft tissues and joint structures, is highly sensitive in detecting synovial cysts that develop in osteoarthritic joints. During ultrasound scanning, the physician longitudinally maneuvers the ultrasound device's scanner while the patient is in a supine position with knees fully extended, facilitating rapid knee imaging assessment. When interpreting the image, the radiologist typically grades osteoarthritis using the KL grade [28]. This evaluation is then utilized by treating physicians or surgeons. A comparison of imaging modalities based on their characteristics is presented in the Table 2.1.

Table 2.1: Key Properties of Imaging Techniques for Knee OA Diagnosis [1].

Imaging Technique		Pros	Cons	Detectable OA features
Radiography		Low cost	Mostly limited to 2D visualization	Joint space narrowing
	Routine OA imaging in clinical practice		Less sensitive to change over time	Osteophyte formation
	Visualising the skeletal structure Scannings in different positions		Lack of soft tissue visualization Prone to positioning errors Risk of radiation	Cyst formation Subchondral sclerosis
MRI	Soft tissue and intra-articular structure visualisation		Expensive	Joint space narrowing
	Cartilage properties and pathology visualization 2D and 3D visualization		Intolerable to metal implant Risk of overdiagnosis	Bone marrow lesions Cyst formation
CT	Bony and calcified tissue visualization		Expensive	Osteophyte formation
	Weight-bearing and kinematic 4D CT		Risk of radiation	Cyst formation
	Allows study of joint metabolism		Contrast injection, may cause allergies Allows 2D (slices) and 3D visualization	Subchondral sclerosis
Ultrasonography		Low cost	Limited to 2D visualization	Osteophyte formation
	Evaluation of ligaments and synovium		Poor contrast	
	Real-time assessment		Knee joint inner margins only	
	Portable Portable		Risk of overdiagnosis Better spatial resolution	
OCT	Real-time evaluation of 4-20 m articular cartilage		N/A for in vivo assessment	Cartilage roughness
	Portable 2D and 3D visualization		Time-consuming	Cartilage degeneration

2.2 Machine Learning Background

In this section, we provide an overview of the machine learning methods utilized in our study. Firstly we explain supervised learning techniques that allow training using labeled data. Subsequently, we explore transfer learning and self-supervised learning(SSL) methods, applicable in scenarios where labeled data is scarce.

2.2.1 Supervised learning

Supervised learning, a subset of machine learning, is characterized by the utilization of labeled data during the learning process by the model [29]. This labeled data is employed as input until appropriate outputs are achieved. Labeled data is essential for constructing machine learning models that exhibit high accuracy or performance. In supervised learning, a training dataset is utilized to derive the desired output from the models. This training dataset encompasses the inputs facilitating the model's gradual learning and their corresponding output pairs. The input data is typically represented by a vector referred to as 'features', and each output is a corresponding 'label' or 'response' to the input. Supervised learning algorithms facilitate pattern recognition within the data, enabling predictions on previously unseen data. The generalization ability of the model, typically relies on a substantial amount of labeled data. The acquisition of labeled data in supervised learning is typically performed by humans, which can render the process time-consuming and costly. Supervised machine learning is generally categorized into two types as regression and classification.

2.2.1.1 Classification

Classification, whose primary purpose is to categorize input data into specific categories, is an important technique widely employed in supervised machine learning. For instance, a classification model can determine whether a tumor is benign or malignant [30]. In this scenario, input data encompasses factors such as tumor size, shape, boundaries, growth rate, patient lifestyle, and family history, while output categories include "benign" or "malignant." Classification models discern patterns and

correlations from training data, enabling accurate classification of data points absent in the training dataset. In deep learning models, input data, which is processed through various layers, is ultimately transformed into a probability distribution by the softmax activation function to generate probability values for each class [31]. The probability distribution generated for each class enables the interpretation for model behavior and evaluating its performance. The ordinal loss function utilized in Chapter 4.2.1.2 is built upon this principle.

2.2.1.2 Regression

Another widely used technique in machine learning is regression, which facilitates the prediction of a continuous output variable by modeling how a dependent variable (output) is influenced by other independent variables (inputs), rather than a predetermined categorical variable. An example of a regression task would be predicting blood pressure based on certain characteristics of patients. A supervised learning algorithm would be trained with data such as patients' age, gender, BMI, and smoking habits, along with corresponding blood pressure values. Once the model is trained, it can be employed to predict mean blood pressure values based on patients' characteristics. The performance of regression models is typically measured using a metric called MAE, as described in Chapter 4.2.2.3.

2.2.2 Transfer learning and self-supervised learning for data scarcity

Developing effective machine learning models without human guidance has long been a recognized challenge. One prominent obstacle lies in the creation of large-scale manually labeled datasets, which are integral for enhancing the accuracy of machine learning algorithms. However, the laborious and expensive process of precise labeling, requiring specialized expertise, leads to a scarcity of labeled training data, which significantly hinders model performance and generalization capabilities. To address this issue, a range of methodologies has been emerged to build advanced deep learning models, thus overcoming the constraints imposed by data scarcity. With transfer learning, particularly in cases where labeled data is limited, the weights obtained through pre-training with a large-scale dataset such as ImageNet can subsequently be

used to downstream tasks with limited labeled data [32]. This approach enhances learning performance by enabling the learned representations of the models to be utilized in new tasks. Self-supervised learning is a subcategory of machine learning. It is developed to facilitate model learning without labeled data [33]. Unlike supervised learning, where models are trained using labels, self-supervised learning entails the model autonomously discerning significant representations within the data. Subsequently, these acquired representations can be refined and applied to downstream tasks, thereby mitigating the necessity for extensive labeled datasets.

2.2.2.1 Contrastive learning

Contrastive learning has arisen as a method of self-supervised learning that mimics the human cognitive process of acquiring new knowledge by contrasting similar and dissimilar entities, rather than learning to recognize entities individually [13]. This technique operates on the premise of instructing similar and dissimilar data points within an unlabeled dataset to the model. The primary objective is to acquire data representations that strengthen the correlation between similar samples, while accentuating the distinctions among dissimilar samples. This is accomplished through the Euclidean distance, which serves to bring similar instances closer together in the representation space while simultaneously pushing dissimilar ones further apart. The process is illustrated in Figure 2.3. The most commonly used loss functions in contrastive learning are contrastive loss and triplet loss.

In the contrastive loss function, a pair of similar or dissimilar samples is provided as input. When the samples are similar, their distance is decreased; conversely, when they are dissimilar, the distance is increased. The formula for the contrastive loss is in Equation 2.1. A pair of samples represented by (I_i, I_j) . The label of each pair denoted by Y . A low-dimensional condensed representation of the features for each sample is provided by CNNs to reside in the embedding space x_i and x_j denote the low-dimensional representations of samples I_i and I_j .

$$L = (1 - Y) * ||x_i - x_j||^2 + Y * \max(0, m - ||x_i - x_j||^2) \quad (2.1)$$

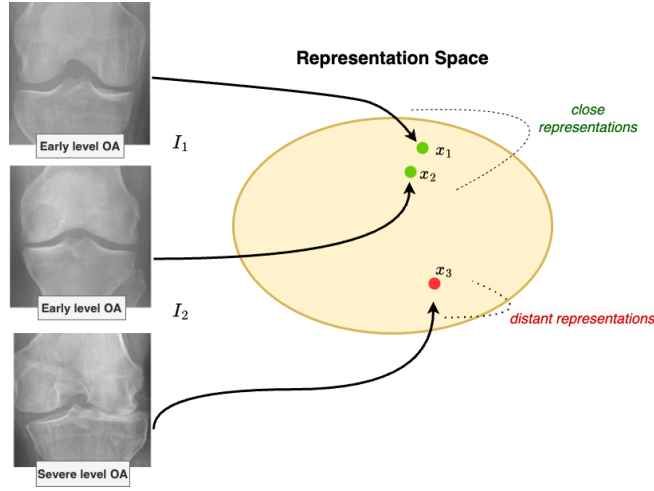


Figure 2.3: Contrastive learning paradigm that bringing similar examples closer, pushing dissimilar ones further apart in representation space.

The parameter m is a positive integer that specifies the minimum distance between instances of different classes. If the samples are similar, $Y = 0$, the term $\|x_i - x_j\|^2$ representing the euclidean distance between the vector representation of two samples is minimised. If the samples are dissimilar, $Y = 1$ and the term $\max(0, m - \|x_i - x_j\|^2)$ is minimised. This is possible by maximising the $\|x_i - x_j\|^2$ term representing the euclidean distance between the two samples.

Triplet loss is a refinement of contrastive loss. It takes three different samples (triplet) representing the reference sample for comparison (anchor), a sample that is similar to the anchor (positive), a sample that is different from the anchor (negative). Equation 2.2 represents triplet loss.

$$L = \max(0, \|x - x^+\|^2 - \|x - x^-\|^2 + m) \quad (2.2)$$

x is the vector representation of the anchor sample, x^- is the vector representation of the negative sample and x^+ is the vector representation of the positive sample. Triplet loss forces that the distance of the anchor sample to the positive sample is shorter than the distance to the negative sample. It is less sensitive to the small number of samples since the weights are updated using both positive and negative samples during model training. This makes triplet loss more effective than contrastive loss.

Contrastive learning is categorized into two types based on whether labeled data is used in the learning process.

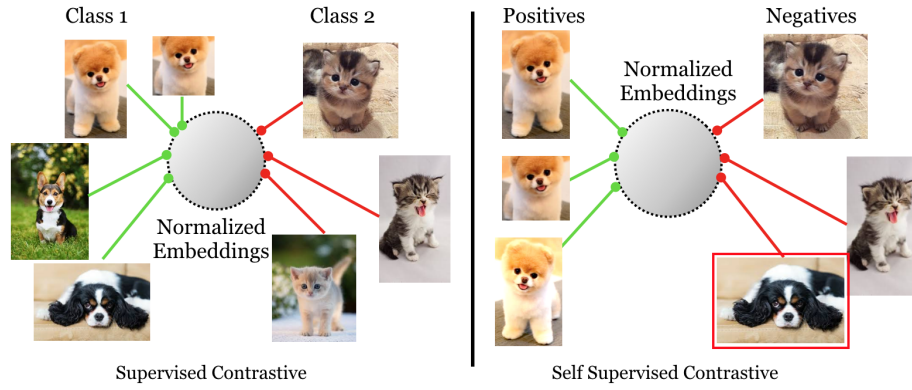


Figure 2.4: Supervised vs. self-supervised contrastive learning [4].

2.2.2.2 Self-supervised contrastive learning

In this approach, the generation of both positive and negative samples relies on the data itself, eliminating the need for manually annotated training data. For instance, positive samples are generated by subjecting the anchor image to random transformations such as jittering, cropping and zooming. The transformations applied to the anchor image should design to yield invariant representations and preserve the basic structure of the data. It often obviates the need for extensive labeled datasets.

2.2.2.3 Supervised contrastive learning

This methodology expands upon the self-checking batch comparative approach by incorporating labeled data within the framework [4]. The positive and negative samples are generated with the provided label information. Samples from same class are strategically positioned in proximity within the embedding space, whereas samples from different classes are separated. It emerged as an alternative when supervised machine learning approaches utilizing standard cross-entropy loss are suboptimal. It is usually used to obtain the encoder in the pre-training phase before fine tuning for downstream tasks, and the frozen encoder is then incorporated into the model during the fine-tuning phase. The difference between supervised contrastive learning and self-supervised learning is simply illustrated in Figure 2.4.

2.2.3 Computer vision

Computer vision is a subset of deep learning dedicated to enabling computers to automatically detect and accurately comprehend images efficiently [34]. Essentially, computer vision applications aim to emulate the human visual system's ability to interpret information. This emulation is achieved using machine learning and deep learning techniques. In this manner, computer vision systems discern patterns within visual data and subsequently seek analogous patterns within other visual datasets, thereby comprehending the visual information. In the workflow of analyzing visual data in computer vision, the process begins with acquiring data through tools like cameras or medical imaging devices. Subsequently, pre-processing steps, such as dimension reduction, noise reduction, and contrast enhancement, are applied to enhance the quality of the acquired images and streamline the analysis process. Following this, deep learning models are typically employed for feature extraction to identify significant visual attributes in the image, such as shape, color, and texture of objects. After feature extraction, recognition and classification of objects in the images are performed. In this step, models are trained using deep learning and machine learning methodologies. Subsequently, post-processing operations are conducted by evaluating and interpreting the obtained results. During this stage, the performance of the trained models is evaluated, and necessary improvements are implemented. Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) are commonly used computer vision models in visual data analysis tasks like image recognition and classification.

2.2.3.1 Convolutional neural networks (CNN)

CNN [35] is a deep learning algorithm frequently used in computer vision tasks, including recognition and classification of images. Due to the different layers within its architecture, it extracts features from the images and subsequently classifies them based on these extracted features. The architecture of the CNN is illustrated in Figure 2.5. The image passes through layers known as convolutional, pooling, and fully connected, undergoing distinct processing within each layer.

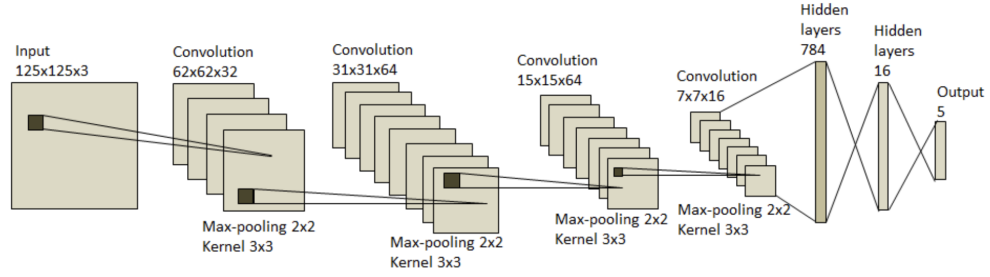


Figure 2.5: An example of CNN architecture [5].

In the convolutional layer, the first layer encountering the input image, the convolution process is applied to extract a feature map containing the image's features. Within each convolutional layer, filters smaller than the original input size are utilized. These filters are shifted across the image to capture specific features. The parameters learned during training in CNN models are the values within these filters. As the model undergoes backpropagation, it updates these values, progressively enhancing its ability to detect image features. In a CNN, an activation function is applied after each convolutional layer, with Rectified Linear Unit(ReLU) being the most commonly used. This function is a simple linear function that eliminates negative inputs and only propagates the positive inputs to the subsequent layers. Pooling layers, another component of CNNs, are employed to decrease computational costs by reducing the dimensions of feature maps. At the end of the CNN architecture, there are fully connected layers. These layers are the conventional neural network. They enable the feature map to become flattened. These layers are capable of performing classification or regression tasks.

2.2.3.2 Vision transformers

The Vision Transformer [6] has emerged as an alternative to CNNs, which are most commonly utilized in image analysis tasks. It has been demonstrated that ViT models can achieve nearly four times better computational efficiency and model performance compared to CNNs. ViT is based on a Transformer architecture. Transformers, a deep learning architecture employed to model long-range connections using attentional mechanisms, initially emerged for solving natural language processing(NLP) problems. Since then, transformer-based architectures have become a standard in NLP tasks. Subsequently, the concept of ViT emerged by

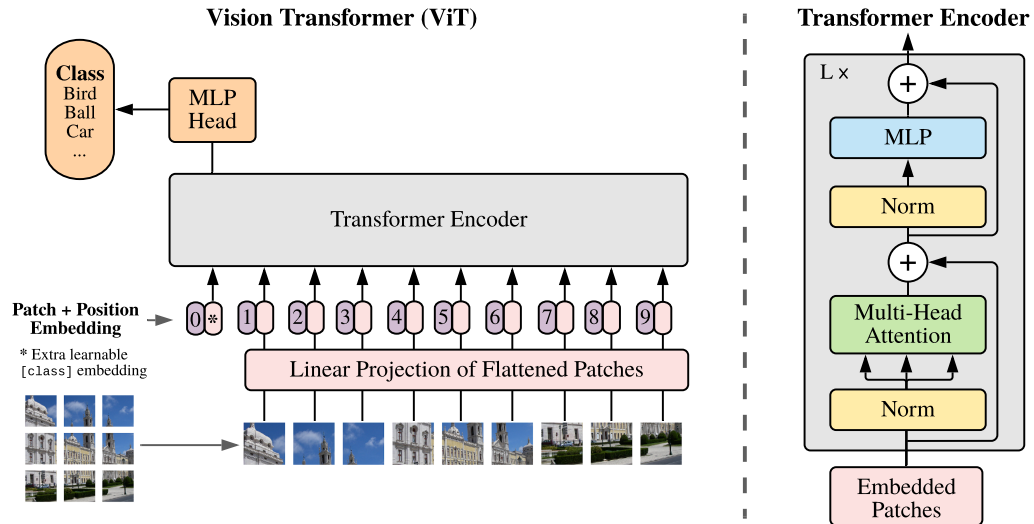


Figure 2.6: Vision transformer(ViT) overview [6].

applying this architecture, originally used in the natural language domain, to the image domain with minor adjustments. Similar to CNNs, ViT comprises different layers designed for various tasks. The model's overview is provided in the Figure 2.6

The input layer partitions the image into smaller segments known as patches, thereby transforming the image into a set of 'tokens' that can be fed into the model. In the positional encoding layer, a code is incorporated to denote the position of each token within the array, facilitating a natural understanding of the relative positions of each token to one another. Transformer encoder layers are the primary components responsible for processing input data. These layers comprise sub-components, including multi-head self-attention mechanisms and feedforward neural networks. These components enable the model to capture intricate relationships and features by discerning dependencies among tokens and facilitating efficient information flow throughout the network, thereby ensuring stabilized training. The classification head is the output layer of the model. It is responsible for classifying the image or performing another task by utilizing the outputs from the attention blocks.

3. LITERATURE REVIEW

This chapter reviews the work of researchers on artificial intelligence approaches in medical fields. As our research focuses on the automated detection of knee OA, we survey the literature on automated diagnosis. Firstly, we examine computer-aided diagnostic (CAD) systems and their usage areas. We list a number of important machine learning and deep learning methods for the automatic diagnosis of knee OA and the classification of disease stages. Lastly, we address the some semi supervised methods employed in disease diagnosis, particularly for the knee OA.

3.1 Computer-Aided Diagnosis in Disease Detection

CAD is the use of computer-generated outputs as an aid by clinicians in diagnosing disease. Traditionally, the diagnosis of diseases relied heavily on the expertise of healthcare professionals, often leading to subjective interpretations and diagnostic errors. The use of CAD is intended to reduce the observational oversight of physicians interpreting medical images and thus reduce false negative rates [36]. It has been used in the radiology for many years, even before advances in artificial intelligence. But the expected results still cannot be achieved with CAD. Therefore, although CAD is not currently replacing medical doctors, it is clear that it is assisting radiologists in the diagnostic process by playing a supporting and interpretive role in medical diagnoses. CAD is most commonly used today to provide quantitative image measurements or to calculate probability values for different diagnoses. [37] CAD systems are widely used in disease diagnosis in various medical specialties, including radiology, oncology, cardiology and pathology.

While in the early days these systems required manual feature engineering based on domain knowledge, today deep learning is often used for the automatic discovery of hidden features in imaging data [38]. It typically works on medical images obtained through a variety of imaging methods, such as radiology, nuclear medicine, MRI, CT,

ultrasonography, and other imaging techniques. For example, with the help of CAD systems, potential tumors for breast cancer detection can be detected on mammography or lesions can be identified on chest radiography [39,40]. The performance of CAD tools varies depending on the characteristics of the images used during system training and the design of the system.

3.2 Machine Learning Techniques for Knee Osteoarthritis

The significance of employing machine learning models for knee osteoarthritis detection has been established in studies by [41] and [42]. In [42] stated that the use of ML models in knee OA studies is a relatively new approach, and underlined that it is a promising development in more accurate grouping of patients. Research on OA has yielded numerous findings crucial for understanding its progression over the years. These encompass findings obtained through imaging, certain substances in blood serum, and the symptoms experienced by patients. They reported that determining the significance of these findings could lead to the identification of different types of OA. Thus, machine learning algorithms were actually used to determine the relative importance of each finding or variable in knee OA detection. Machine learning algorithms trained using specific biomarkers that identify osteoarthritis help identify individuals at high and low risk [42].

In [43], showed that innovative ML approaches are very useful in clustering knee OA into predefined groups. They employed publicly accessible data from the Foundation for the National Institutes of Health, spanning 48 months, and applied Distance Weighted Discrimination. The primary biomarkers influencing disease progression at 48 months were identified as ,bone spurs, bone marrow lesions, medial meniscal extrusion, and urine C-terminal crosslinked telopeptide. Conversely, biomarkers such as WOMAC pain,lateral meniscal extrusion and serum N-terminal pro-peptide of collagen IIA (PIIANP) were found to have negligible impact. With this study evaluated many markers of different types and scales simultaneously and identified those most relevant to the progression phenotype.

In [44], authors developed a novel image-based CAD software with knee X-ray images and machine learning algorithms to classify as healthy or OA. They utilized

a collection of 1024 knee X-ray images sourced from the publicly available Osteoarthritis Initiative (OAI) database. The method they propose consists of 5 stages. First of all, they extracted 3 different ROIs from a knee radiograph. They are medial, middle and lateral sides. Then, these 3 extracted parts are preprocessed in the Fourier domain with the circular filtering method to ensure the stability of the images. In order to preserve essential information, a quantization of the gray level is applied on the image. To reduce the variability between samples, they applied intensity normalization using multivariate linear regression (MLR) to data. Independent component analysis (ICA) is then used to minimize the dimension of the normalized ROIs. Thus, each sample could now be represented by a certain number of features. The random forest and Naive Bayes classifiers were used to classify the X-ray images according to these features extracted from the image. With 82.98% accuracy and 87.15% sensitivity scores, their proposed approach outperformed many methods in the literature.

In another study in 2017 [45], the authors presented a method that could not only detect knee osteoarthritis, but also produce as output the distinctive parameters of how this prediction is made, using random forest. The authors state that this fills the gap in confidence in automated computer systems by ensuring the explainability of decision processes, in contrast to other studies in the literature. In the study, random forest regressors were used to analyze ground reaction forces to identify knee OA patients. They collected motion data from 84 individuals, including 47 patients diagnosed with OA and 47 healthy participants. These individuals walked on two paths equipped with three-component force sensors. Various parameters related to the anterior-posterior and medio-lateral ground reaction forces, such as average values, push time, and slope, were recorded. Subsequently, random forest regression models were employed to correlate these parameters with the severity of knee osteoarthritis. To enhance the model's generalization capability, a 5-fold cross-validation approach was implemented.

3.3 Deep Learning Approaches for Predicting Knee Osteoarthritis

Nowadays, deep learning models have gained rapid popularity owing to the burgeoning volume of data and advancements in computational power. Furthermore, the

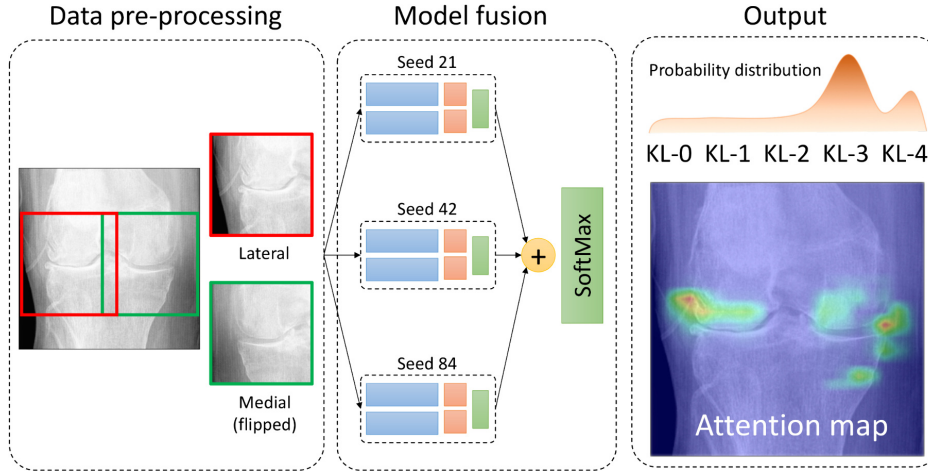


Figure 3.1: Proposed classification pipeline in [7] study.

advancement of imaging technologies has underscored the significance of deep learning models within the medical sphere. Recently, it has become the most important component of many systems, including automatic knee OA detection systems.

In a study [7], the authors proposed a CAD system that classify the grade of knee OA automatically according to the KL scale. They utilized a Deep Siamese CNN in the study. To demonstrate the robustness of their method, they employed distinct datasets for both the training and testing phases. They propose a new method to use Deep Siamese CNN architecture for joint part on knee X-ray using cropping and symmetry. As seen in Figure 3.1, they crop the right and left sides of the joint area from the knee X-ray image and they flipped the medial side. So they provide two inputs to the model. Thus, they aimed to make the network more resistant to noise by reducing the number of parameters. The authors utilized the Multicenter Osteoarthritis (MOST) dataset for method development and the Osteoarthritis Initiative (OAI) dataset for validation. Furthermore, they employed an ensemble structure, training three different models initialized using different random seeds. So they simulated the process of utilizing multiple interpreters when labeling datasets in real-world scenarios. Subsequently, for each seed, they selected the model exhibiting the optimal performance on the validation dataset. At the end, by providing probability values for each KL grade, they are enhancing transparency in physicians' decision-making processes. Consequently, they achieved an accuracy score of 66.71% for radiological OA diagnosis. It was the highest reached score in the literature to date.

In 2016, authors in [46] study proposed a novel method to classify knee OA stages from radiographs with CNNs. This paper highlighted that an evaluation metric based on continuous distance such as mean square error, is more suitable metric for classifying OA severity compared to classification accuracy. Thus, the task of quantifying disease grades, traditionally approached as a classification problem, was reframed as a regression problem, employing regression loss during model training. The study reported an improvement in model performance when utilizing the regression loss function.

In [47], the authors developed a multi-task deep learning model for the identification of patients who have undergone total knee replacement (TKR). This model utilized the knee radiographs, specifically based on posteroanterior (PA) view tibial femoral (TF) joint radiographs. This model is capable of predicting the probability of TKR within a span of 9 years and the KL grade from the radiographs. They employed ResNet-34 along with seven-fold nested cross-validation incorporating transfer learning, achieving an AUC score of 0.87.

In another study [48], YOLO v2 detection model used the for knee joint detection to extract of knee joint regions containing important information for knee OA detection. They have introduced an ordinal loss function to classify knee OA grades based on Kellgren-Lawrence. This function designed to strongly penalize significant discrepancies between predicted and actual KL grades. They then fine-tuned frequently used CNN models and obtained a score of 69.60% with VGG-19.

In the study [49], a novel representation model is introduced to identify early stages knee OA, leveraging auto-encoders. A penalty term is incorporated into the cost function to enhance the differentiation between OA and healthy subjects. Primarily, the study involves the extraction of five distinct regions of interest (ROIs), each sized 32x32, from knee radiographs, capturing the lateral, central, and medial portions of the knee. The proposed model extract features from each ROI, which are subsequently utilized for class separation. The experimental findings highlight the medial part is the most distinguishing region on the knee radiograph for distinguishing between the two groups.

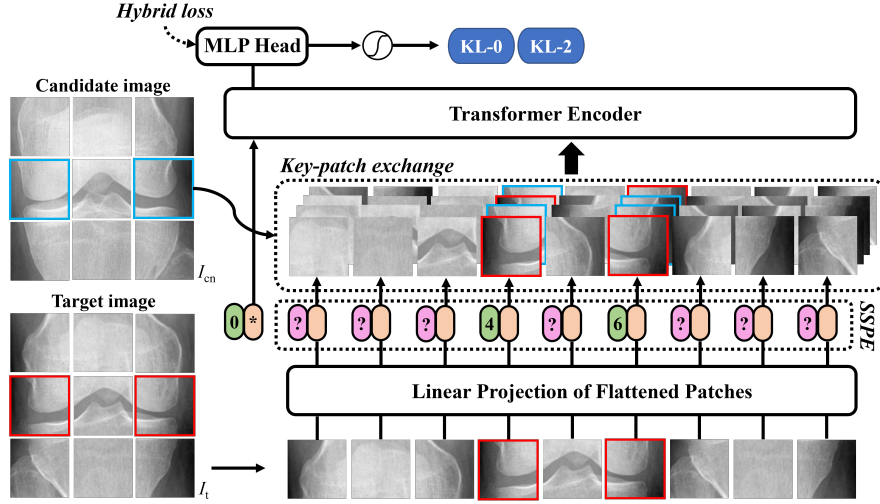


Figure 3.2: Flowchart of proposed key-patch exchange strategy in [8].

In recent times, vision transformers have emerged as an alternative approach to the CNN-based methods traditionally predominant in the field of image processing. They have demonstrated remarkable efficacy, yielding impressive results when trained on large-scale visual datasets [50]. This shift is attributed to the enhanced scalability and generalization capabilities offered by vision transformers compared to their CNN-based counterparts. Vision transformers excel in capturing long-range relationships and processing image information across various scales. Consequently, their adoption is steadily increasing in the medical domain, where they contribute to advancements in disease diagnosis and imaging applications. One of these studies is a transformer-based framework was introduced in [51]. This framework simulates the interaction between the treating physician and a radiologist. The system comprises a radiology module, composed of a CNN and a transformer, for feature extraction, and a general practitioner module that integrates these features with clinical data. The authors reported the superior performance of their proposed approach in contrast to other techniques documented in the literature.

In [8], a novel data augmentation technique was introduced for the early detection of knee OA utilizing the ViT model incorporating original selective shuffled position embedding (SSPE). The approach focused exclusively on KL-0 and KL-2 stages, as KL-0 exhibits no signs of osteoarthritis while knee OA is definitively present at KL-2, albeit with minimal severity. Through the ViT-based learning model and a hybrid loss

function integrating the key patch replacement strategy, the accuracy achieved in this investigation surpassed that of the CNNs commonly used in the literature. In such a way, SSPE can force ViT model to concentrate solely on the critical patches relevant to OA. Proposed key-patch exchange strategy illustrated in Figure 3.2.

3.4 Semi-Supervised Learning and Contrastive Learning Based Methods

Deep learning models' success largely relies on manually labeled data. Under normal circumstances, labeling data is a time-consuming process; in the medical field, it is even more challenging because it requires expertise and can only be performed by experienced doctors. This presents a significant obstacle to developing high performance models, which in turn limits clinical applications. To address this issue, in the literature, there are novel methods for disease detection that utilize both labeled and unlabeled data and provide good model performance even if the number of labeled data is small.

For example, in [52], authors proposed a semi-supervised framework using labeled and unlabeled knee 3D MRI scans to detect knee OA. In this framework, instead of classic CNN models as feature extractors, they designed a 3D deep CNN model called 3D-ResCNN-GAP. As the name suggests, this model consists of 3D CNN blocks, Global Average Pooling (GAP) layers, and residual CNN. Within the SSL approach, they trained the 3D-ResCNN-GAP model using labeled data and then made inferences for unlabeled data using this model. Subsequently, they treated these inferences as pseudo-labels, combined labeled and unlabeled data, and re-trained the 3D-ResCNN-GAP model. By incorporating unlabeled data into the framework in this manner, they effectively augmented the labeled training data. The impact of employing the SSL framework was investigated across different knee MRI acquisitions, enhancing model performance across all acquisitions.

In [53], a semi-supervised framework for the detection of diabetic retinopathy from fundus images has been proposed. Initially, in the pre-training , self-supervised contrastive learning was performed using unlabeled data. Then, a model was fine-tuned using a small labeled dataset, leveraging weights obtained from the pre-training phase, to classify fundus images as either referable or non-referable diabetic retinopathy

(DR). In the contrastive learning part of the pre-training phase, a neural style transfer (NST) module was applied as the data augmentation technique. This is the key contribution of the study. The paper indicates that natural transformations such as rotation and flipping do not alter the texture of the image, and therefore contribute less to representation learning. However, with NST, the texture of the input image is altered while maintaining its semantic meaning, thereby facilitating better representation learning. The entire pipeline has been named FundusNET. The EyePACS dataset was used for training and validation, while clinical data collected from the University of Illinois was used for testing. The original SimCLR pipeline and a lesion-based CL method were used as baseline models. FundusNET demonstrated superiority over both baseline models.

In another study [54], the ResNet50 model was fine-tuned to detect the presence of Covid-19 from chest X-ray images. The initial weights of the ResNet50 model were obtained from ten different models that had been pre-trained using either supervised or self-supervised approaches. This was done to investigate which pre-training method had a more positive impact on the final results. Using weights from the self-supervised method Swapping Assignments Between Views(SwAV) provided excellent results on the validation data. The lower performance of the model pre-trained with a supervised approach on another chest dataset proves the superior effectiveness of using self-supervised methods in pre-training.

In another study related to the SSL approach [55], a method called dual semi-supervised learning (DSL) is presented for classifying the severity of Alzheimer's disease using a limited number of labeled neuropsychological test data. First, from 64 items across 7 neuropsychological tests, those significantly affecting Alzheimer's detection are identified using the Pearson correlation coefficient, and the top 15 items (features) with the highest coefficients are selected. Then, by using these 15 features as input in the SLL framework, the disease stages are identified. In the SLL framework, both labeled and unlabeled data are used. Initially, features of each labeled data point are extracted by two different encoders, and cross entropy loss values are obtained by comparing these with the true labels. Subsequently, these encoders are applied to the unlabeled data. Two separate representations are obtained

from the encoders using the unlabeled data, and cross entropy loss is calculated based on the probability values of the encoders relative to each other. Throughout this framework, efforts are made to minimize the loss value. The performance of the model is measured using various metrics with a total of 60 and 120 labeled data points in the SLL framework. Accuracy values of 85.47% with 60 labeled data points and 88.40% with 120 labeled data points were achieved. It has shown superiority when compared with different semi-supervised methods.

Changes in bone surface morphology over time can lead to the development of knee OA. Therefore, in [56], the authors propose a deep neural network to predict bone shape changes 48 months in advance to elucidate the disease's progression. In this study, we see a different use of contrastive learning other than semi-supervised. Using three MRI scans over 36 months from the OAI dataset, the authors predicted the spherical encoded maps of the bone shape 48 months ($T_0 + 48m = T_4$) after the last scan using contrastive learning. First, the femur, tibia, and patella bone regions were segmented from the MRI scans using a bone segmentation model trained with 57 MRI images. This segmentation model is based on the V-Net encoder-decoder architecture. The segmented masks were used to isolate the bone shape, which was then encoded using 2D spherical maps. A ResNet-18 was then trained to stage the disease according to KL grades using these spherical maps. To predict the change between the spherical map at the ($T_0 + 24m = T_2$) and 36 months later ($T_0 + 36m = T_3$), a deep iterative down-Up network was trained. A new loss function was designed for this model, consisting of the sum of the structural similarity score and the contrastive loss value between the actual and predicted changes. By adding the predicted change amount to the spherical map at T_2 , the bone shape at T_3 was predicted. Features of this predicted bone shape were extracted using ResNet-18. The contrastive loss was obtained using these features and the features of the real bone shape at T_3 . Using this pipeline, the bone shape at T_4 was predicted.



4. SUPERVISED LEARNING METHODS IN KNEE OA SEVERITY CLASSIFICATION

In this chapter, we described our fully supervised methods for classifying knee OA stages using labeled knee X-ray images. The architectures utilized, loss functions employed, selected hyperparameters, details regarding the data, and methodologies are presented within the separate sections. Finally, we presented the results and analyze each in detail.

4.1 Model Architectures

Various deep learning architectures exhibit distinct performances contingent upon the dataset and problem domain, owing to their unique feature extraction capabilities and learning capacities. This study aims to assess the performance of models constructed with different deep learning architectures in the classification of knee osteoarthritis severity. This section outlines the model architectures employed in our research.

Firstly, we used ResNet, one of the most popular architecture in image recognition and classification. It has provided a significant innovation in CNNs to increase the depth and performance of the network. While it has been theoretically proven that deep networks provide more accurate results, the accuracy of models consisting of twenty five layers decreases as the layers increase, that is contrary to this [9]. Because as the models get deeper, the optimization of the model becomes more difficult due to vanished gradients. For this reason, an architecture was introduced called ResNet to solve the vanishing gradient problem. [57]. Due to residual connections in this architecture, the model is able to learn the difference between input and output by adding its input directly to the output of the next layer, rather than the output of a typical layer. Hence, the back-propagation pathway flows through the residual link, maintaining a gradient multiplication factor of one. This mechanism serves to retain information and mitigate the risk of gradient vanishing. Further insights into the configuration of a residual block are depicted in Figure 4.1, that illustrate how the

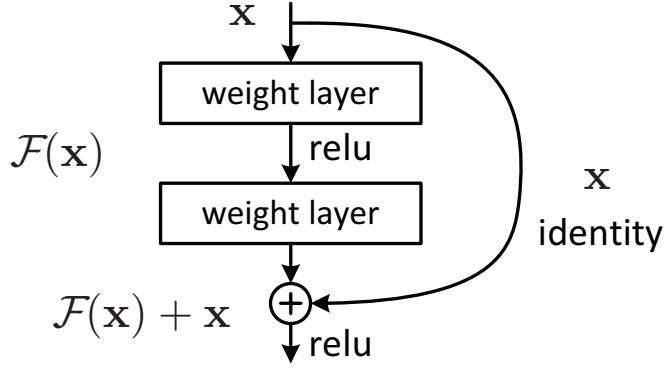


Figure 4.1: Residual block [9]. Skip-connections or residual-connections that is a solution for vanishing gradient in deep networks.

identity matrix supplements the output layer prior to entering the ReLU layer. ResNet is formed by the combination of multiple residual blocks. These blocks generally enable the creation of ResNet models with different depths. In our study, we used the ResNet-152 version, which consists of 152 layers.

ResNeXt, one of the architectures we used, is an architecture that enhances the learning capacity of neural network and improves generalization ability by utilizing parallel group layers and cardinality connections to increase the scalability of the model [58]. Parallel structure increases the scalability of the model. This provides better performance on larger and more complex datasets. ResNeXt is based on the ResNet architecture but differs from ResNet in that it uses "groups" instead of many small paths. These groups contain multiple parallel paths with filters to learn different features. Thus, the network learns more features more effectively and its representation ability increases. ResNeXt consists of building blocks called cardinality groups. The term cardinality parameter refers to the number of parallel paths in the group. Parallel paths in each group are used to capture different features from the input data. We used ResNext-50 in our study.

The DenseNet architecture offers a different approach to mitigate the vanishing gradient problem. With this approach, the propagation of features between layers becomes easier and the parameter count has been significantly decreased. [59]. In standard network architectures, every neuron within one layer is linked with all neurons in the subsequent layer. This means that features from previous layers are transmitted

to subsequent layers through a single path. In the DenseNet architecture, unlike a standard neural network, every neuron within a layer is linked to every neuron in all the preceding layers. This allows each neuron to share features more effectively by receiving output from each of the neurons in the previous layers. For classification tasks, the DenseNet-201 model has proven to be more successful than other DenseNet models [60]. For this reason, we preferred DenseNet-201 in our study.

While increasing the size and complexity of models for solving difficult tasks increases accuracy, it also increases the processing load of the models [10]. As a solution to this, the efficient and scalable EfficientNet architecture has been introduced to reduce both training and inference times [61]. With this architecture, the neural network is scaled not only in depth, width and resolution, but also in all three. Depth indicates the number of channels present in the layers. The model can capture more complex patterns as the number of channels rises. Depth is related to the number of layers in the network. While increasing the number of layers enables more complex patterns to be captured, it also increases computational costs. Resolution is related to the input size. Higher resolution images are more informative for the model, but again increase the computational cost. The effect of the width, depth, and resolution variables on the model efficiency is demonstrated in Figure 4.2. EfficientNet discovers the best balance between performance and resource consumption by finding the optimal scaling constants for these three key dimensions with composite scaling. At the same time, the mobile inverted bottleneck layers in its design, combined with squeeze-and-excitation optimization, make the neural network more efficient. We used EfficientNet-B0 and EfficientNet-B7, variants of the EfficientNet architecture with different scaling constants.

One of the model architectures we used is the ConvNeXt architecture. During the 2020s, ViTs emerged as the preferred option for generic vision backbones, surpassing CNNs [11]. However, a belief has emerged that ViTs are more accurate and scalable than ConvNets. Thereupon, ConvNeXt, a pure ConvNet model that can compete with ViTs and re-demonstrate the importance of convolution, has been proposed. ConvNeXt is an evolution of the ResNet architecture with different training procedures and innovative architecture. The training procedures and changes in the architecture

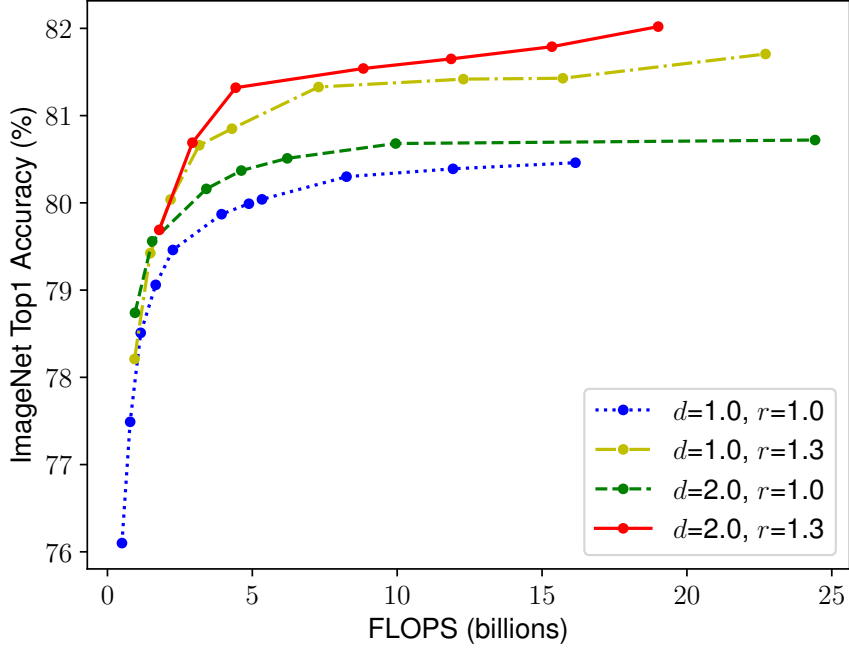


Figure 4.2: Effect of width, depth and resolution on model efficiency. The points on the lines represent different models with different width coefficient (w) [10].

were inspired by the methods used in Swin Transformer, one of the state of the art (SOTA) ViT models.

Removing the pooling layer by applying convolutional filters of size (4,4) to an input image in (4,4) step size, having fewer normalization layers and replacing the BatchNorm2d layer with a LayerNorm2d layer that can operate independently of the batch size are some of the innovative features of ConvNeXt architectures that improve performance. ViTs analyze images by leveraging self-attention mechanisms and partitioning the image into smaller units known as patches. ConvNeXt is inspired by ViTs on these topics. However, it does not use directly the same technique called as transformer in ViTs. ConvNeXt uses a different method called depthwise convolution. With this method, different parts of the image are processed separately. This technique reduces the computational cost while maintaining the performance of the model. However, ConvNext-V1 has a feature collapse problem. This is identified by the existence of redundant activations (inactive or overloaded neurons) throughout the channels and negatively affects the feature extraction of the model. In order to increase the feature extraction ability in ConvNeXt architectures, a higher version,

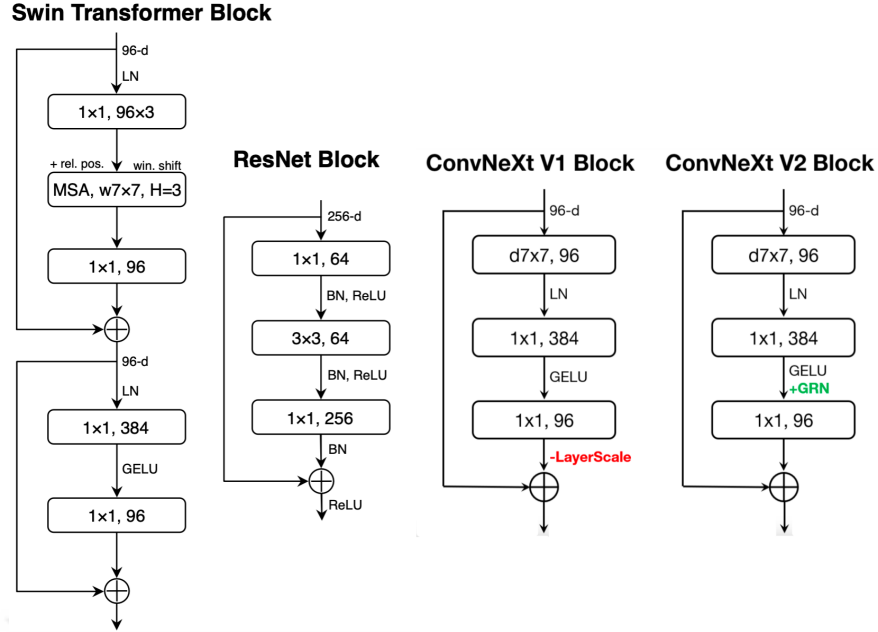


Figure 4.3: Block designs for a ResNet, a Swin Transformer, a ConvNeXt V1 and a ConvNeXt V2 [11,12].

ConvNeXt-V2 architecture, has emerged with the addition of fully convolutional masked autoencoder structures and new global response normalization (GRN) layers [12]. Placing the GRN layer following the dimension-expansion MLP layer alleviated the feature collapse problem. GRN increases the contrast and selectivity of channels. Because an input feature goes through three stages of the GRN unit. These stages are global feature collection, feature normalization and feature calibration. Block designs of ConvNeXt-V1 and its advanced version ConvNeXt-V2 models, inspired by Swin Transformer and ResNet models, are illustrated in the Figure 4.3.

4.2 Loss Functions and Evaluation Metrics

In this section, we provide information about the loss functions we used when training the models and the evaluation metrics we used to evaluate our models. Loss functions are used as objective functions in forward propagation and backpropagation processes that make the training of deep learning models possible. During forward propagation, a loss value is calculated by comparing the final output obtained after the computational operations on the input images with the actual outputs. Our main goal is to reduce this loss value by updating the learnable weights through the optimization function in the backpropagation phase.

4.2.1 Loss function

The choice of loss function, which is used as an important guide in optimising the model, is critical to the accuracy of deep learning models. In our study, we employed the cross entropy loss function, which is used as a standard for classifier models. We also employed the ordinal loss function. It applies a penalty by taking into account the distance difference between the actual class and the predicted class in problems where there is a ranking between the labels.

4.2.1.1 Cross entropy loss function

In image classification problems with more than two classes, the cross entropy loss function is commonly used. Cross entropy is calculated by the formula given in Equation 4.1.

$$L(\hat{y}, y) = - \sum_{k=0}^{n-1} y^k \log(\hat{y}^{(k)}) \quad (4.1)$$

In formula, n represents the number of classes. If k is the correct class, $y^{(k)}$ takes the value 1, if k is the wrong class, it takes the value 0. $\hat{y}^{(k)}$ represents the estimated probability value for class k . In this scenario, the greater the estimated probability value for the actual class, the smaller the loss. The estimated probability values for the other classes are not taken into account in this calculation. Thus, all misclassifications are penalised in the same way. However, since the labels are ordinal, the distance between the true class and the predicted class must also be taken into account by the loss function. For example, when the model predicts an image with the real label KL-0 as KL-4, it should not be expressed with the same loss value as KL-1. Cross entropy loss assigns a loss value to all classes equally in case of error predictions and does not take into account the proximity between classes, which leads to underperforming models.

4.2.1.2 Ordinal loss function

Ordinal loss is a type of loss function used in classification problems for ordinal labels. There is an order between labels (e.g. from small to large or from few to many) and these labels cannot be considered independent of each other. Due to the ordinal nature

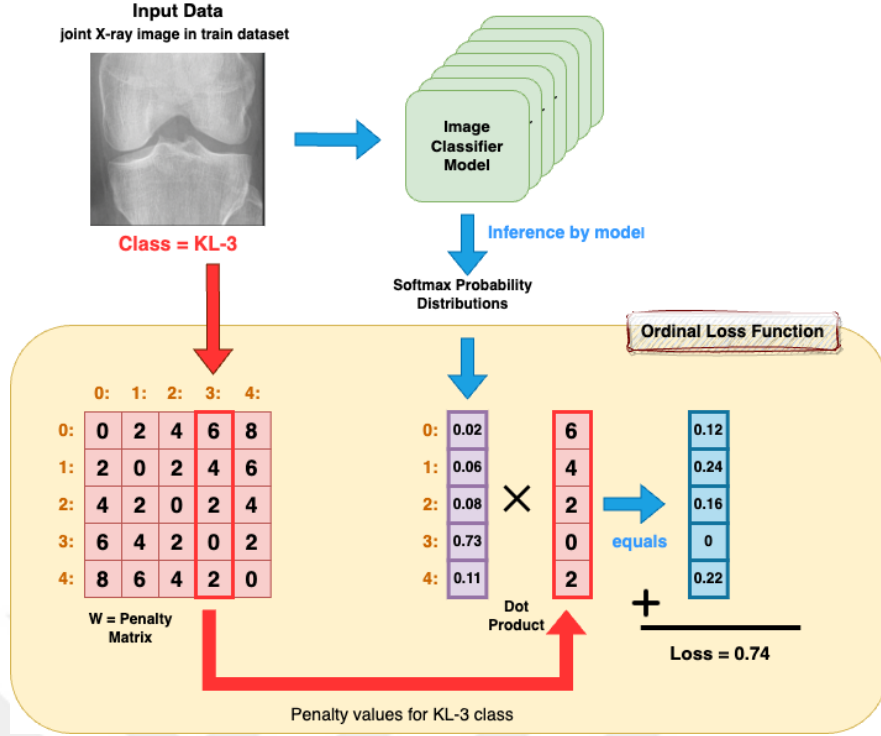


Figure 4.4: Ordinal loss calculation schema.

of the labels in our dataset, we integrate the ordinal loss function proposed in the [48] to ensure that the models we develop to classify the severity of knee OA also take into account the distance between the real classes and the forecasted classes. Thus, the model does not ignore the magnitude of the error during fine-tuning and learns the relationship between the classes better.

$$L(\hat{y}, y) = \sum_{i=0}^{n-1} w_{i,m} \cdot p_i \quad (4.2)$$

The equation for the ordinal loss function we apply is given in Equation 4.2. In formula, W represents a square matrix of size $n * n$. It serves for defining penalty weights with respect to the true KL class. $w_{i,m} \in W$ denotes the penalty weight for predicting class m as class i and p_i denotes the probability value obtained from the softmax layer for class i . For the KL grade classification task, $n = 5$ and $i, m \in \{0, 1, \dots, n-1\}$. Figure 4.4 shows the process of calculating the loss value with the ordinal loss function using the W matrix containing the penalty weights and the probability values obtained by the softmax layer.

4.2.2 Evaluation metrics

As machine learning and deep learning models have become tools used in every field today, accurately evaluating the performance of the models has become a very important step. Because evaluation makes it possible to comprehend how well the model works with real world data and enhancing of models. Especially in medical applications, this is of vital importance because ensuring the performance of models is necessary for accurate diagnosis of diseases and effective treatment plans. Therefore, a detailed study and evaluation of model performance in medical studies is a fundamental step to improve the quality of healthcare services and so improve the quality of life of patients. In order to provide reliable results in this respect, we use several metrics that allow us to analyze the performance of our classifier models from different perspectives.

4.2.2.1 Weighted accuracy

It is a metric often used to measure the accuracy of a model in classification problems. Mathematically, it is the proportion of accurately categorised samples to all samples. Related formula is in Equation 4.3. TP represents to the situation where a model correctly classifies, TN refers to the situation where the model correctly classifies as negative, FP denoted to the situation where the model classifies an example that is actually negative as positive, and FN is the model incorrectly classifies an example that is actually positive as negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.3)$$

Since the knee OA data used in our study has a ordinal structure and an unbalanced distribution between classes, weighted accuracy alone is not a sufficient metric to assess the performance of our models.

4.2.2.2 F1 score

F1 score is a performance evaluation metric calculated by the harmonic mean of precision and recall. It used to measure the quality of classifier models. It takes a

value between 0 and 1, indicating that the performance of the model improves as it approaches 1.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4.4)$$

The precision value is the number of true positives divided by the number of all positives so it shows the percentage of positive results identified correctly.

$$Precision = \frac{TP}{TP + FP} \quad (4.5)$$

Recall value is calculated by dividing the the number of true positives by sum of number of true positives and false negatives. So recall measures the rate at which true positives are correctly detected.

$$Recall = \frac{TP}{TP + FN} \quad (4.6)$$

Recall reveals a model's ability to avoid missing true positive outcomes. Since early detection of diseases is critical, a large recall value is desirable in this case. The F1 score can provide more insight into the quality of the model than accuracy, particularly in cases when the distribution of classes is not balanced. The models we developed have two main purposes. These are facilitating early diagnosis and swiftly detecting advanced KL grades that require surgical intervention. These objectives necessitate accurate prediction for each class. Therefore, in our study, we employ the macro-F1 score variation, which assumes equal importance for all classes. The equation for the macro-F1 score is given in Equation 4.7. In the equation, i represents KL grades. The F1-scores calculated for each class are summed and then divided by the number of classes represented by n .

$$MacroF1\ Score = \frac{\sum_{i=0}^{n-1} F1\ Score_i}{n} \quad (4.7)$$

4.2.2.3 MAE

Mean average error (MAE) is a performance measure normally used for regression problems. It is the average of the absolute values of the differences between the true values and the predicted values. It is used to evaluate how accurate a prediction model is. For n observations, MAE calculated with the formula given in Equation 4.8.

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (4.8)$$

y_i is denoted the real value and $\hat{y}_i$ is denoted the predicted value. In the identification of stages of knee OA, the occurrence of misclassifications between distant stages is more critical than those between adjacent stages. Consequently, relying solely on classification accuracy and F1-score metrics may prove inadequate in evaluating this aspect. Taking inspiration from the use of MAE as a performance measure in age estimation problems [62], we utilized the MAE as our performance metric.

4.2.2.4 sa-RPS

In classifier models, the probability distribution obtained through the softmax layer, generally determine model accuracy and model reliability. Especially in disease diagnosis prediction or disease stage classification, these probability values produced by the model for each class are informative clues about the model. For example, the prediction of "healthy" with a probability of 49% by a model that classifying sick or healthy is a warning for clinicians and encourages them to request further tests. This is the transparency expected in the medical field. So, it is sometimes not enough for the model to predict correctly, and in this case it is expected to predict correctly with a high probability value. There are metrics that are popular in the literature, such as Quadratic Weighted Kappa and Expected Cost [63,64] to assess the performance of ordinal classification models. However, a recent study [65] has shown that they are not optimal solutions for assessing the quality of the probability distribution and sa-RPS has been proposed to address this problem.

In our study, we used the proposed sa-RPS metric to evaluate the performance of our classifier models. We also observe the effect of the approaches we apply to our

classifier models and examine how close these approaches bring us to the optimal model. The formula for this metric is in Equation 4.9.

$$sa-RPS(p, y) = \frac{1}{K-1} \left[\sum_{i=1}^K \left| \sum_{j=1}^i (p_j - y_j) \right| \right]^2 \quad (4.9)$$

K is the number of classes. The probability of the model is denoted by p_j and the true label is denoted by y_j . Figure 4.5 shows the probability distributions given by two different models on the same X-ray image based on our 5-class KL grade classification problem. Taking these distributions into account, the process of calculating the sa-RPS metric for both models is shown. The probabilities of the first model are denoted by p_1 and the probabilities of the second model are denoted by p_2 . P_1 of the probabilistic prediction p_1 and P_2 of the probabilistic prediction p_2 . The true KL degree value is KL-2, and considering the probability distributions, both models are correct. However, when we look at the distribution of the first model, the probabilities of KL-1, KL-2 and KL-3 classes are close to each other and there is a symmetrical situation. In the second model, there is no symmetrical situation and the model seems to be undecided between KL-1 and KL-2. It is preferable for KL grade classification if the model falls between 2 grades that are close to each other and gives high probabilities to these 2 grades. However, being indecisive among the three classes is a worse scenario and results in a higher sa-RPS score.

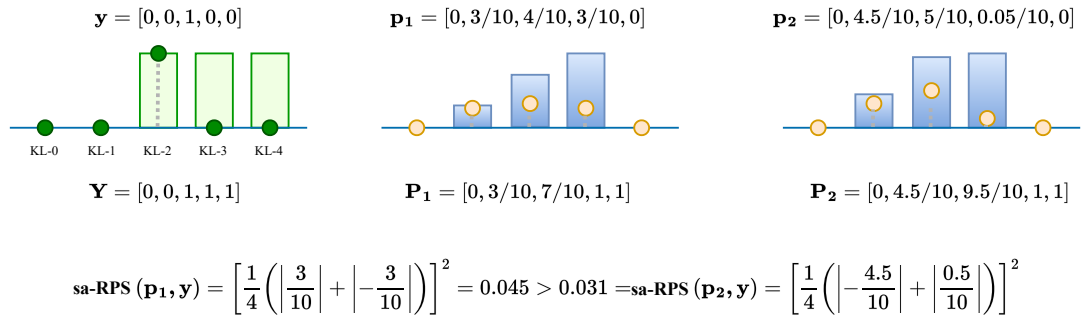


Figure 4.5: Calculation process of sa-RPS score.

4.3 Experimental Setups

We trained our models on NVIDIA Quadro RTX 6000 GPU using the Keras library with version 2.11.0.

4.3.1 Parameter setup

Hyperparameters are parameters that are not learned during the training process and are determined by the user before starting model training. They may have a direct impact on deep learning models' performance. One of these hyperparameters is the learning rate, which represents the step size when updating the model parameters during backpropagation. While the learning rate may be a constant value throughout learning, there are also approaches that allow it to increase or decrease during learning. We also used this approach in our study. The learning rate, which was initially 0.0001, decreases by 0.2 when there is no decrease in the loss value for 10 epochs. Thus the model is better able to understand the overall structure of the data without getting bogged down in more details.

Another hyperparameter that greatly affects the training speed and generalization ability of the model is the optimization algorithm. These algorithms allow parameters to be updated to minimize the loss of the model. Although Adam optimizer [66] is the most widely used optimisation algorithm, it is observed that it is inadequate when the number of parameters is large. Rectified Adam (RAdam) optimizer [67] has emerged to address the weaknesses of Adam and aim to perform better on large-scale models. Initially, RAdam behaves like normal Adam optimisation, but as training progresses, the gradient variance decreases and a correction mechanism is introduced to ensure the accuracy of the momentum. So, we used RAdam optimizer to provide more stable training in our large-scale models.

Another critical hyperparameter is the number of epochs. It reflects the number of times entire training dataset is shown to the network during training, and hence how long the model is trained. Selecting the correct number of epochs is crucial for model's capability to perform effectively on data beyond the training set. If the model's epoch number is too low, this results in a situation called underfitting that is, the model can't learn sufficiently. On the other hand, if the epoch size is too large, it triggers a situation called overfitting. In this case, a higher accuracy score obtained on the training dataset but fail on the test dataset it has never encountered. To avoid these situations, it is important to monitor both the training and validation performance of the model

throughout the training period. Based on this, the early stopping mechanism [68] automatically stops the training in cases where the performance in the validation data set does not increase or the loss value increases. We also applied the early stopping approach in our study in order to establish a balance in terms of performance between the training and test sets if the model performance does not improve in six epochs.

4.3.2 Datasets

We used two different datasets in supervised learning methods part of our study. We firstly used publicly accessible OAI data for the one stage fine tuning part. OAI [69] is a research program focused on osteoarthritis research, and for this purpose, researchers have created a community to follow individuals at risk of osteoarthritis between the ages of 45 and 79. This dataset consists of raw knee X-ray images and their corresponding KL grades. Due to the multi-centric nature of OAI, the image resolution and dimensions of the knee X-ray images are not consistent. Therefore, there is a need for resizing the raw X-ray images to ensure they have the same resolution. However, prior to that, it is necessary to extract the joint region from the raw X-ray image of the entire knee. This is because the joint region is the focus area for the identification of knee OA. So, we used a dataset named as Knee Osteoarthritis Severity Grading Dataset [70], which is already created by extracting the focal joint regions for the right and left knee from the raw knee radiography images on OAI. In this study, the detection of knee joint regions was performed using the YOLO-V2 algorithm and as a result, knee joint regions were extracted. There are a total of 8260 samples in this dataset, which is splitted into train, validation and test sets in a proportion of 7: 1: 2.

The second dataset we used in our study is our in-house dataset. It was collected from the Radiology department of Sivas Cumhuriyet University Hospital between 2020-2023. This dataset includes a total of 877 samples, consisting of knee X-ray images of 564 female and 313 male individuals aged between 14 and 92 years. The knee X-ray images in the dataset were annotated according to their KL grade by two experienced radiologists and a rheumatologist. In this dataset, we maintained the proportion of 7:1:2 while separating it into training, validation and test sets. The data distribution between classes for both datasets is in Table 4.1. The numbers in the table

Table 4.1: Data distributions among classes and subsets for the public OAI dataset and the in-house hospital dataset.

Grade	OAI			Hospital Data		
	Train	Validation	Test	Train	Validation	Test
Grade 0	2286	328	639	150	22	43
Grade 1	1046	153	296	166	24	47
Grade 2	1516	212	447	125	18	36
Grade 3	757	106	223	114	17	32
Grade 4	173	27	51	58	8	17
Total	5778	826	1656	613	89	175

depicted how many knee X-rays were used for each KL class. Subjects in the train, validation and tests sets were different for both dataset. The models were trained using the training data, while the validation data were utilized to adjust the hyperparameters and assess overfitting. The test set was employed to estimate the final performance.

4.3.3 Detail implementation

In this section, details about the pre-processing steps applied to the data before training and training workflows will be included.

4.3.4 Data preprocessing

Within this section, we cover the procedures applied to raw data with the aim of preparing it before entering the model. Firstly, for the hospital dataset used in the second stage of our two-stage fine tuning implementation, knee X-ray images were extracted from the data in DICOM format. Then, the joint areas of these images were manually extracted by the clinicians. Due to manualization, images are not a fixed size. For deep learning models expecting fixed input size, so all images are resized to 224x224. Since the pixel values of the images are between 0 and 255 in both datasets, these values are scaled between 0 and 1. Then, standardization techniques such as samplewise center and samplewise standardization are used to normalize the statistical properties of the data in both datasets. Each sample of the data is centered by subtracting its own mean and standardized by dividing it by its own standard deviation.

Finally, random transformations are applied to the data in training set to help the model learn in a more general and robust way for both datasets. These are saturation, contrast, brightness and angle transformations applied randomly on the images. Such transformations allow the model to experience a wider range of variation during the training process, thus reducing overfitting and increasing its ability to generalize. After all pre-processing steps, the data can be seen in the Figure 4.6.

4.3.4.1 One-stage fine-tuning

In this part of our experiments, we investigate the success of an approach that incorporates traditional deep neural network architectures as well as newer architectures in the automatic detection of knee osteoarthritis severity from knee joint radiographs. In addition, we observe the result of applying the adjustable ordinal loss function on the learning of the model and reveal the effect it has on the model performance. Our method of implementing deep learning model architectures is fine-tuning models that pre-trained with the large-scale ImageNet dataset [71] using knee X-ray images and their associated KL labels. During the training process, we train the model with the training data mentioned in Chapter 4.3.2 and obtain the performance of the model on the validation data at each epoch. We perform all our experiments on the OAI dataset. Before training, the data preprocessing steps mentioned in Chapter 4.3.4 were applied.

4.3.4.2 Two-stage fine-tuning

Our two-stage approach essentially describes the approach we proposed for in-house hospital data. We assumed that the models in the first stage learned key features related to the diagnosis of knee OA and the classification of the severity of knee OA from a larger number and variety of OAI data. For this reason, utilizing the models acquired in the initial stage, we examined the impact of the features extracted by these models for OA diagnosis from the OAI data on a hospital dataset as a novel dataset. Consequently, we elucidated approaches to enhance model performance for the new hospital dataset. In this stage, we fine-tuned the models in different architecture from the first stage, using a hospital dataset that smaller size and collected from a single centre. The models

we fine-tuned at this stage are ConvNeXt-Base Ordinal, EfficientNet-B7-Ordinal and DenseNet-201-Ordinal, which are among the successful model architectures of the first stage. The batch size for all experiments was 16 and the maximum epoch size was 50 with early stopping approach.

We conducted each experiment three times for every model, calculating the performance metrics by averaging the results across these three trials. Additionally, to investigate the impact of re-fine tuning models initially fine-tuned with OAI data, we also fine-tuned the ImageNet pre-trained versions of the same model architectures using hospital data. During the first stage of training, we unfroze all layers; however, in the second stage, we froze all layers except the last 100. This approach aimed to retain important features learned from the initial model while facilitating the acquisition of new hospital data insights through the last 100 layers. We used the grid search algorithm to determine how many layers to freeze initially. Unfreezing the last 100 layers yielded the most favorable outcomes across all three models. In the second stage of two stage fine tuning method, we investigated whether the phenomenon observed in the first stage, employing ordinal loss in the OAI dataset, would manifest similarly in the hospital dataset. Hence, we utilized both ordinal loss and cross-entropy loss functions during model training. The two-stage fine-tuning schema is depicted in Figure 4.6.

4.4 Results and Analysis for Supervised Learning Methods

In the one-stage fine-tuning process, we trained supervised classification models employing various architectures using OAI data to predict the knee OA stage. We employed both the categorical cross-entropy loss function and the adjustable ordinal loss function. Subsequently, in the two-stage fine-tuning process, we retrained the models obtained from the one-stage fine-tuning using in-house hospital data, maintaining the utilization of the same loss functions. Moreover, we fine-tuned models that pre-trained models with ImageNet data using in-house hospital data. This made it easier to assess the two-stage fine-tuning approach's effectiveness. In this section, the experimental results are examined under three distinct headings.

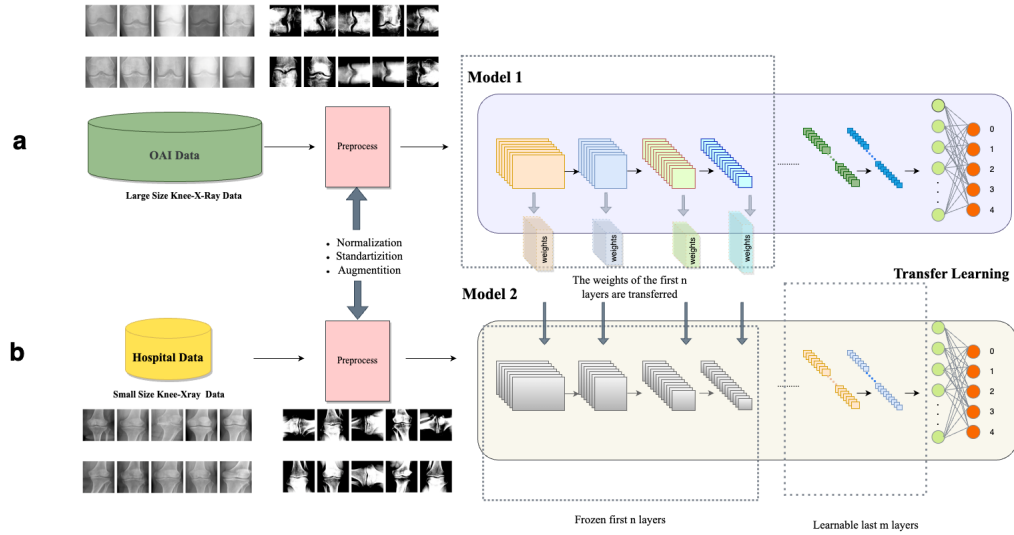


Figure 4.6: In the two-stage fine-tuning schema, fine-tuning stages are seen for 2 different datasets. In stage 1(a), a pre-trained model with ImageNet weights is fine-tuned with larger data from OAI dataset, and the weights of this model are transferred to the model in stage 2(b) and the all layers are frozen except the last 100 layers. In stage 2, it was fine-tuned again with less data from our in-house dataset, using the same pre-processing steps and the same model architecture as in the first stage.

These are model architecture, utilized loss function, and the application of the two-staged fine-tuning method on in-house hospital data.

Table 4.2: The performances of the models that were fine-tuned once using the OAI dataset.

Models	Loss Function	Accuracy(%)	F1-Score(%)	MAE
ResNet-152	Cross entropy	69.87	67.30	0.37
	Ordinal	69.38	69.58	0.35
DenseNet-201	Cross entropy	69.44	69.15	0.37
	Ordinal	69.20	70.13	0.34
EfficientNet-B0	Cross entropy	68.23	67.18	0.38
	Ordinal	69.14	69.61	0.35
EfficientNet-B7	Cross Entropy	69.27	68.23	0.36
	Ordinal	71.14	71.25	0.32
ConvNeXt V1-Base	Cross Entropy	72.22	71.84	0.30
	Ordinal	73.12	72.97	0.28
ConvNeXt V2-Tiny	Cross Entropy	73.24	73.16	0.30
	Ordinal	73.91	73.77	0.29
ConvNeXt V2-Nano	Cross Entropy	72.46	71.98	0.31
	Ordinal	73.36	73.51	0.30
ResNeXt50	Cross Entropy	69.14	66.28	0.37
	Ordinal	71.24	70.04	0.31

Table 4.3: The performances of the models that were fine-tuned once using the in-house hospital dataset.

Models	Loss Function	Accuracy(%)	F1-Score(%)	MAE	msa-RPS
EfficientNet-B7	Cross Entropy	61.90	64.17	0.45	0.051
	Ordinal	63.04	65.39	0.47	0.048
DenseNet-201	Cross Entropy	60.76	62.29	0.47	0.046
	Ordinal	61.33	61.56	0.43	0.042
ConvNeXt-Base	Cross Entropy	66.28	68.79	0.44	0.039
	Ordinal	67.80	69.35	0.41	0.036

Table 4.4: The performances of the models that were two-stage fine-tuned using the in-house hospital dataset.

Models	Loss Function	Accuracy(%)	F1-Score(%)	MAE	msa-RPS
EfficientNet-B7	Cross Entropy	63.42	65.40	0.38	0.043
	Ordinal	65.33	67.25	0.35	0.041
DenseNet-201	Cross Entropy	62.28	63.99	0.44	0.046
	Ordinal	62.66	64.43	0.42	0.042
ConvNeXt-Base	Cross Entropy	68.00	69.88	0.36	0.034
	Ordinal	70.86	71.35	0.30	0.029

Table 4.5: Comparison of two-stage fine-tuned and one-stage fine-tuned ConvNeXt-Base model using in-house hospital dataset.

Structure	Model	Accuracy(%)	F1-Score(%)	MAE	msa-RPS
ImageNet	ConvNeXt-Base	67.80	69.35	0.41	0.036
Two stage	ConvNeXt-Base	70.86	71.35	0.30	0.029

4.4.1 Computer vision model architecture comparison

The results of the first stage experiments are presented in Table 4.2. These results were obtained via fine-tuning pre-trained models on ImageNet, using the OAI X-ray images and represents the performance metrics' scores on test set. The ConvNeXt architecture exhibited superior performance when compared with alternative model architectures for classifying knee OA stages. The ConvNeXt V2-Tiny model achieved the highest accuracy and F1 score metrics with 73.91% and 73.77% rates, respectively. In addition, the ConvNeXt V1-Base model demonstrates superior performance, evidenced by MAE value of 0.28. Subsequently, ConvNeXt V2-Nano, a small-scale model comprising 15.6M parameters, exhibits performance on par with these two models. Additionally, the approximate parity in performance between the ConvNeXt V2-Tiny model, which boasts 28 million parameters, and the ConvNeXt V1-Base model with 89 million parameters, underscores once again the efficacy of enhancing inter-channel feature competition within ConvNeXt-V2 architectures.

Regarding model performance, both the EfficientNet and ResNeXt architectures closely approximate the performance of ConvNeXt models. However, models within the ConvNeXt architecture exhibit significant superiority for both loss functions compared to other models. Despite having fewer parameters, the ConvNeXt-v2-Tiny model outperforms larger models in other architectures. This circumstance also highlights the effect of the depthwise separable convolution method in ConvNeXt models. With this technique, the number of parameters is reduced by employing separate filters for each channel, followed by cross-channel interactions through channel-wise information fusion. Furthermore, compared to other architectures, the smaller filter sizes in the convolutional layers facilitate better capturing of minimal changes in the joint located within a small pixel area in X-ray images. Consequently,

lighter models with fewer parameters demonstrate more accurate classification of disease stages.

The results of one-stage fine-tuned the models Imagenet pre-trained with Imagenet using in-house data and the two-stage fine-tuned models on in-house data are given respectively in Table 4.3 and Table 4.4. In both tables, the performance superiority between the three models is once again attributed to the ConvNeXt models. This observation shows that ConvNeXt models' performance superiority across various datasets. To the best of our knowledge, ConvNeXt architectures have not been utilized before in the literature for assessing the knee OA grading.

The success of the ConvNeXt-Base model over the ResNet-152 model is further evidenced by attention maps generated using the Grad-CAM [72] technique. This is a visualization technique used to understand which regions are taken into account in the classification tasks of deep learning models and was developed to explain the model's predictions. Figure 4.8 displays an X-ray image of a knee joint with a screw inserted in the test dataset. To predict the KL grade of this knee, we employed the ResNet-152 and ConvNeXt-Base models that we trained in the first fine-tuning stage. In the ResNet-152 model, the presence of the inserted screw led to incorrect predictions. However, the ConvNeXt-Base model disregarded this unexpected object in the image and successfully detected the KL-2 phase by focusing on the inter-articular space.

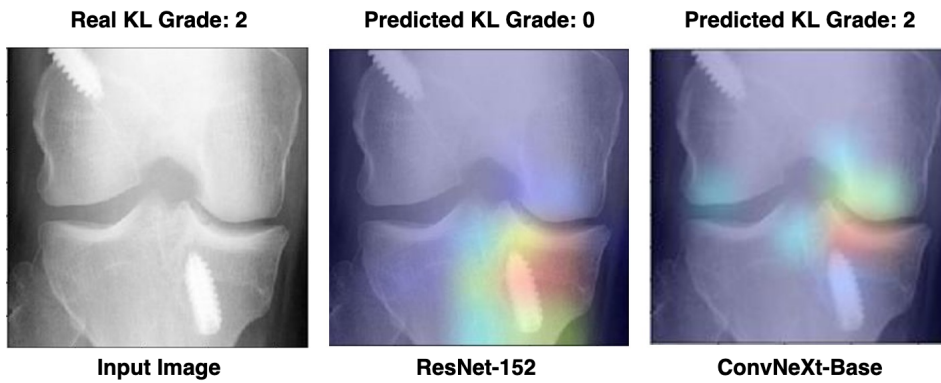


Figure 4.7: Attention map comparison using Grad-CAM technique.

4.4.2 Loss functions comparison

Table 4.2 illustrates the results derived from the first stage fine-tuning. Herein, we observed the classification performance of models fine-tuned using knee X-ray images with both cross-entropy and ordinal loss functions. These results were acquired through the fine-tuning of pre-trained models with ImageNet data on OAI X-ray images. The performance scores were derived from the OAI test set. Models trained using ordinal loss and cross-entropy loss functions in different architectures facilitate a fair comparison of these two loss functions. Because, independently of the model, we can see the effect of loss functions on model performance. When comparing models trained with cross-entropy and ordinal loss functions, the superior performance is observed the models trained with ordinal loss.

In the versions trained with ordinal loss across all architectures, higher accuracy scores and F1-scores along with lower MAE values are generally achieved compared to versions trained with cross entropy loss. As an exception, slight decreases in accuracy are observed in the versions of ResNet-152 and DenseNet-201 models trained with ordinal loss. However, in the same models, significant decreases in MAE metrics and increases in F1 scores demonstrate that misclassifications are not among distant classes but rather among closer classes, indicating an improvement in model performance. Moreover, in this study focusing on classifying KL stages, it also underscores the unreliability of assessing model performances solely through the accuracy metric. In Table 4.4, we present the outcomes of our proposed two-stage fine-tuning method on in-house hospital data. Table 4.3, on the other hand, illustrates the performance on the in-house hospital test data achieved by fine-tuning models pre-trained on ImageNet with our in-house hospital data. In these fine-tunings, both ordinal loss and cross-entropy loss were used. We have examined whether the performance improvement observed with ordinal loss on OAI data in Table 4.2 is also evident in our in-house hospital data. According to the results obtained from these two tables, it is observed that the use of ordinal loss in both one-stage and two-stage fine-tuning leads to performance improvements across all three model architectures. Through the

mean sa-RPS metric utilized in evaluating our classifier models, we once again observe the beneficial impact of ordinal loss during training. This demonstrates that sa-RPS can effectively capture the positive impact of ordinal loss on model performance, making it an appropriate metric for our problem domain. Consequently, it presents an opportunity to assess our models from a different standpoint.

In conclusion, based on our experimental results, independent of model architecture, methodology employed, and dataset utilized, we provide evidence through various metrics. The ordinal loss function is more suitable for classifying the severity levels of OA disease compared to the commonly used cross entropy loss function.

4.4.3 Paradigm for fine-tuning comparison

With the two-stage fine tuning method, we first investigated the effect of the features extracted from OAI data by the models on the performance of a new dataset. Thus, we are investigating methods that will improve model performance for the in-house hospital dataset. For this purpose, we utilized the models obtained from the first stage for the second stage of fine-tuning. The models we used from the first-stage are EfficientNet-B7, ConvNeXt V1-Base, and DenseNet-201. We chose these models because they have different architectures, and we aimed to examine the effect of two-stage fine-tuning across different architectures. Table 4.4 shows the scores of our classifier models on the in-house hospital test data using the two-stage method.

The hospital training data contains 613 knee X-ray images. This number is not sufficient for extracting significant features for classifying knee OA degrees. We illustrated this by the poor performance of the fine-tuned models using only hospital training data in Table 4.3. These results are the scores of models pre-trained with ImageNet and fine-tuned with in-house hospital training data on the test set of hospital data. So we can say that two-stage fine-tuning boosts the performance of all three model architectures. Figure 4.8 shows the Grad-CAM results of the ConvNeXt-Base model the best model which we obtained with two-stage fine-tuning. When we look at these activation maps created by the Grad-CAM method, we see that our high-performance model generally makes accurate predictions by focusing on the pixels from joint space narrow and osteophyte.

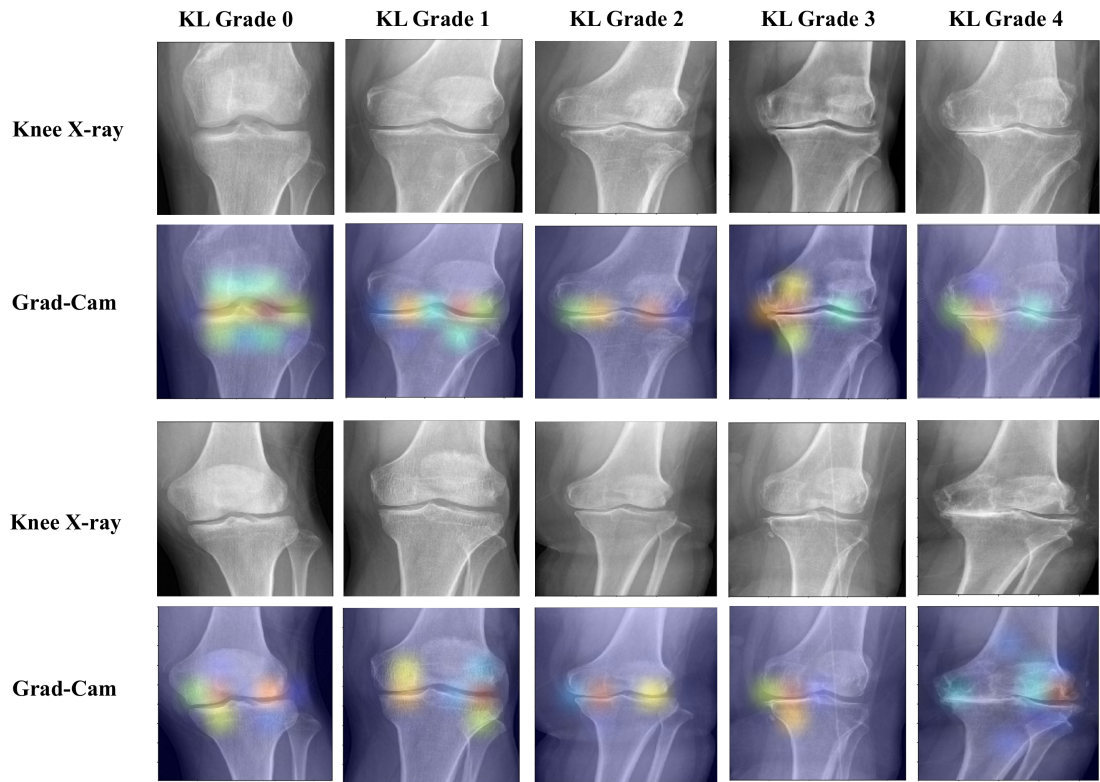


Figure 4.8: Localizations produced using the Grad-CAM method on the ConvNeXt-Base model. This model was obtained by a two-stage fine-tuning strategy and ordinal loss on hospital dataset. The first and third row are samples of knee X-rays and the second and fourth row are their corresponding overlaid classification activation maps. Joint space narrowing and osteophyte formation were successfully captured when the model predicted the grade of knee OA.



5. SELF- SUPERVISED LEARNING METHODS IN KNEE OA SEVERITY CLASSIFICATION

Data labeling is a labor-intensive process that often requires a significant amount of effort. Consequently, training models without human supervision has long been a persistent issue. Finding labeled data for many tasks can be challenging. Particularly, labeling medical images is a specialized task that can only be done by a limited number of experts. Additionally, diagnosing some diseases can be more difficult, further complicating the labeling process. This situation presents the greatest obstacle to creating large-scale datasets and limits model training. Therefore, self-supervised learning approaches that eliminate the dependency on labeled data and allow training with unlabeled data have gained popularity recently. One of these approaches is contrastive learning, which we discussed in Chapter 2.2.2.1. We used contrastive learning to learn from unlabeled data. In this section, we provide details on how we employ contrastive learning and present our experimental results.

5.1 Proposed Semi-Supervised Paradigm

In supervised learning, successful image classifiers often rely on the presence of large amounts of labeled data. However, in real-world scenarios, limited amounts of labeled data are often available, while accessing unlabeled data is much easier. Taking this scenario into account, we employ SimCLR, one of the most popular contrastive learning techniques, to enhance our model performance in situations when we have limited amount of labeled data [13]. SimCLR aligns representations of augmented views of a data sample in the latent space by applying various transformations to the data, using a contrastive loss function. This framework is depicted in Figure 5.1. The SimCLR includes a data augmentation module. With this module, a positive pair is created by applying predefined random augmentations to a data sample. These augmentations are denoted as x_i and x_j . Each of these augmented pairs subsequently passes through a neural network base encoder, denoted as $f(\cdot)$, for extraction of

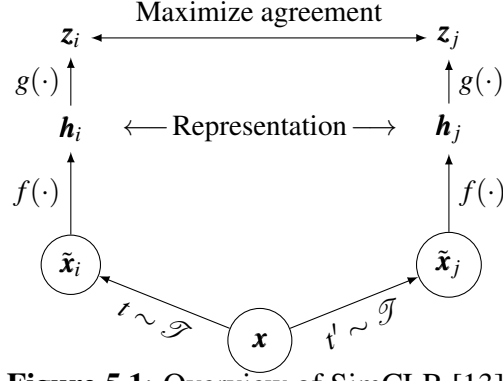


Figure 5.1: Overview of SimCLR [13].

representation vectors denoted as h_i and h_j . Later, a small neural network, referred to as the projection head $g(\cdot)$, is utilized to represent feature vectors or embeddings extracted in the previous stage in a more suitable space. Subsequently, contrastive loss is employed to reduce the distance between these vectors generated, aiming to bring similar examples closer to each other in the space. Positive pairs are generated for all images in mini-batch. In a mini-batch consisting of N samples, all samples except the selected one are considered as negative examples, amounting to $2(N - 1)$ in total. Figure 5.2 summarizing our proposed semi-supervised learning approach, utilizing unlabeled knee X-ray images and a limited number of labeled knee X-ray data. Our proposed framework comprises two primary steps.

Firstly, we conduct self-supervised pretraining on unlabeled X-ray images using SimCLR to acquire visual representations. Following the training of the encoder model in this stage, it is then transferred to the second step for supervised fine-tuning on labeled X-ray images. The weights transferred from the pretraining stage are learned from unlabeled data via contrastive learning. Visual representations are learned as the model attempts to distinguish between similar and dissimilar samples (Figure 5.2a). In this part, by using contrastive loss on the positive examples generated from input image, the weights of the encoder and projection head updated. This process helps bring images with similar KL grades closer together in the representation space. During the supervised fine-tuning stage (Figure 5.2b), the model initiates fine-tuning with these weights on a limited number of labeled data.

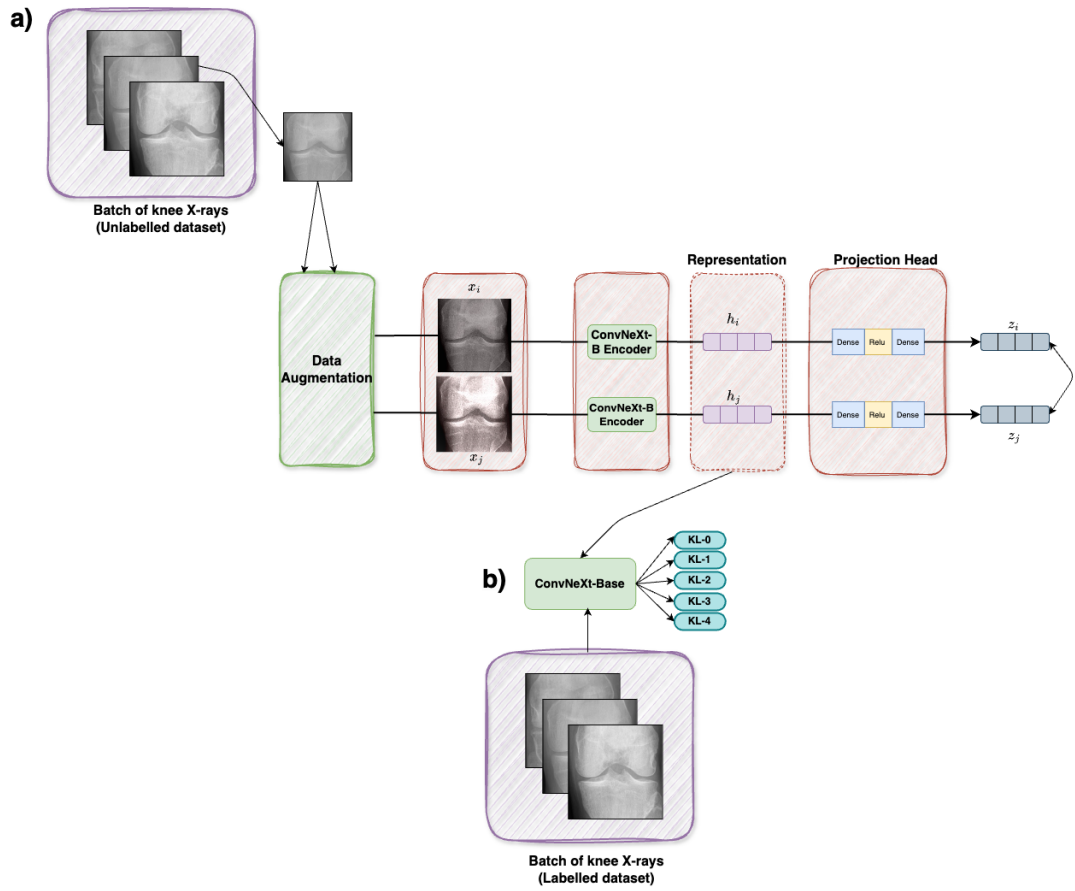


Figure 5.2: Proposed semi-supervised framework based on contrastive learning.

5.2 Experimental Setups

We trained our models on NVIDIA Quadro RTX 6000 GPU using the Keras library with version 2.11.0. The training time of the pretraining model lasted 2 hours and 56 minutes.

5.2.1 Parameter setup

One of the most crucial parameters for contrastive learning is the batch size. This is because, within the SimCLR framework, while taking a random example within the batch as input and generating its positive samples, all remaining examples are considered as negatives. However, some of these negatives might belong to the same class as the input image. In such cases, these data instances from the same class will be pushed apart in the latent space by the contrastive loss, which will hinder the learning process. Conversely, having a higher number of negative samples from different classes within a batch is essential, as it ensures that they are distant in the latent space. When the batch size is large, the probability of encountering true negative examples with different labels increases; therefore, batch size is an important parameter. For the semi-supervised framework we propose, during the pre-training phase from unlabeled data, our batch size is 32. In the supervised classification part, however, the batch size is 16.

Another important hyperparameter is the learning rate. During the pre-training phase, the initial learning rate is set to $0.5e-3$. While this is an appropriate value initially for the learning process from unlabeled data, it is necessary for the learning rate to decrease over time to enhance the generalization ability of the model. To achieve this gradual decrease in the learning rate, the cosine decay function is employed. This method aids in stabilizing the model by nonlinearly reducing the learning rate over time.

In contrastive learning, another important hyperparameter is the temperature parameter. While the contrastive loss determines the similarity values between positive and negative examples using the Euclidean distance, these values are scaled by a

temperature parameter. A lower temperature value increases the punishment for negative examples, thus prompting the model to pay more attention to harder negative examples. We experimented with values such as 0.1, 0.05, and 0.2 for temperature. As a result, we achieved the lowest loss during the pre-training stage by 0.1.

5.2.2 Datasets

For the proposed semi-supervised framework to forecast the KL grade of OA, we used OAI dataset for all our experiments. We partitioned this dataset into different proportions for our experiments. In dataset, there are a total of 5778 labeled knee X-ray images. Initially, we divided these images into two subsets, consisting of 4777 and 1001 images, respectively. While doing this, we did not disrupt the data distribution between classes. The subset of 4777 images is utilized for pre-training the encoder via contrastive learning in an unsupervised manner. In order to examine the label efficiency of our proposed framework, we trained our classifier model in the second stage on different partitions of 1001 labeled training data, and subsequently tested each trained model on the test dataset. Gradually reducing the labeled dataset, employing 100%, 75%, 50%, and 25% of labeled data successively, we trained the classifier model. We keep the validation set and test set within the OAI dataset unchanged, utilizing them consistently to test our supervised fine-tuned model each time. When dividing the training dataset into four different proportions, we ensure that the inter-class data distribution remains consistent. The inter-class data distribution for the training dataset created at various proportions in the Table 5.1

5.2.3 Data preprocessing

In the SimCLR framework and other contrastive learning techniques, the importance of transformation operations applied to input data to ensure the invariance of input data during the generation of positive examples has been demonstrated many times in the success of encoder models. So, strong data augmentation should be used carefully. The reason for using augmented versions of input images is to ensure that these transformations don't cause changes in the labels of the input image, thus allowing the model to learn important features belonging to the input image. When predicting

Table 5.1: Data distribution between classes in progressively reduced training data for supervised fine-tuning stage.

Grade	Validation Data	Test Data	Train Data			
			100%	75%	50%	25%
Grade 0	328	639	396	297	198	99
Grade 1	153	296	181	135	90	45
Grade 2	212	447	263	197	132	66
Grade 3	106	223	98	66	33	99
Grade 4	27	51	30	22	15	7
Total	826	1656	1001	749	501	250

disease stages in knee X-ray images, the most important information lies in the pixels containing osteophytes around the knee joint and the joint space. Therefore, we do not apply random crop, which is commonly used as strong data augmentation in the SimCLR framework, because we do not want these pixels from the input image to be lost. The transformations we apply include color jittering, rotation, Gaussian noise, Gaussian blur, and flipping. Therefore, during the pretraining stage, we use these transformations, while during the supervised fine-tuning stage, we use those described in Chapter 4.3.4.

5.2.4 Detail implementation

We iterate the framework outlined in the Figure 5.2 from begin to end three times, including the pretraining phase. Our experimental results are derived by averaging the scores obtained from these three repetitions for each performance metric. During the pretraining phase, the initial weights of the ConvNeXt V1-Base integrated as the encoder are randomly assigned. As described in Chapter 5.1, we initially employ the SimCLR method for pretraining with 4777 unlabeled data, utilizing the ConvNeXt V1-Base architecture to construct an encoder. We then explore the impact of features extracted from a large volume of unlabeled X-rays on the performance of the classification model by supervised fine-tuning, using a small labeled X-ray dataset comprising 1001 samples. Hence, with the weights in this encoder, we initiate the fine-tuning process. To better demonstrate the impact of the encoder, we also utilize a supervised baseline based on the ConvNeXt V1-Base architecture. We fine-tune

Table 5.2: Performance comparisons between our proposed SimCLR-based model and the baseline model on the OAI test dataset.

Train Data Rate	Pretraining	Accuracy	F1-Score	MAE
100% data	SimCLR	64.16	64.72	0.42
	Baseline	62.53	63.05	0.47
75% data	SimCLR	61.27	62.65	0.46
	Baseline	56.42	57.05	0.56
50% data	SimCLR	57.48	56.83	0.51
	Baseline	49.36	49.91	0.65
25% data	SimCLR	52.13	51.44	0.60
	Baseline	41.33	40.10	0.88

this supervised baseline using the same labeled datasets, leveraging weights from pretraining on ImageNet. Iteratively, we repeat the supervised fine-tuning stage using each of the four datasets we created by gradually reducing the number of labeled data. We conduct this process both with and without utilizing the encoder. Subsequently, we record the performance on the same test dataset after each iteration.

5.3 Results and Analysis for Semi- supervised Learning Methods

The scores of supervised classification models fine-tuned using four different datasets, created by progressively reducing the amount of data, are provided in the Table 5.2. A summary of the results in the table is also depicted in Figure 5.3. Results vary depending on whether the proposed pretraining approach or baseline models pretrained with Imagenet weights were utilized. The best performance was achieved with the model fine-tuned using CL pretraining on the entire training dataset (1001 data points). On the other hand, as the number of data decreased, the gap between the baseline model and model with proposed method increased. The pretrained CL models exhibited good performance even with a 75% reduction in training data. A ConvNeXt V1-Base network pre-trained with contrastive learning demonstrates the profound impact of self-supervised learning for developing more robust classification models with less data.

While transfer learning via ImageNet can be beneficial in solving many problems, recent studies have underscored the potential limitations of features learned from ImageNet for medical imaging tasks. Because it consists of natural images unrelated

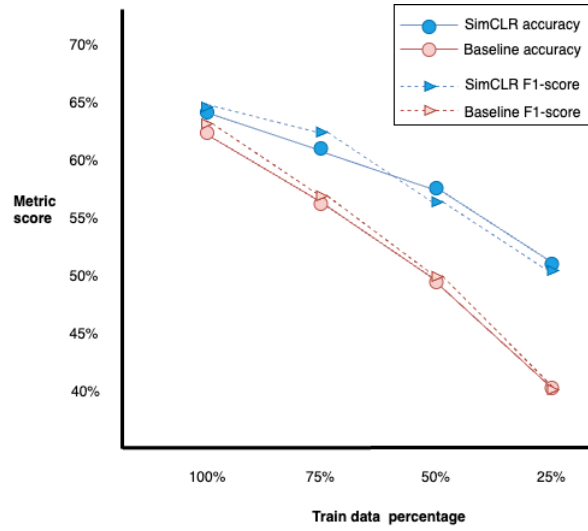


Figure 5.3: Accuracy and F1-score for our proposed SimCLR based model and baseline model on test set.

to medical images. Therefore, we emphasize that our results highlight the potential for using medical images in pretraining to be much more effective than relying on ImageNet.

6. CONCLUSIONS

In this section, firstly we explained our approach to answering the main questions related to our research. We then elaborated on the contributions of this thesis. Finally, we discussed limitations and future work related to our research.

6.1 Answers

Throughout our research process, we addressed the emerging questions by employing various methodological approaches. Below, we provide the answers to the questions raised in **Chapter 1**:

- **Question 1** How do different deep learning architectures perform in fully supervised learning?

Our answer to this question is described in **Section 4.4.1**, where we compare eight different architectures on the OAI dataset and three different architectures on the in-house hospital dataset. Among the architectures we selected, the ConvNeXt V1-Base and ConvNeXt V2-Tiny models gave the best results.

- **Question 2** How much a deep model is sensitive to choosing the right loss function for the severity of knee OA classification task?

In **Section 4.4.2** we compare two loss functions in the models trained to classify the stages of knee OA. We employ the categorical cross entropy loss function, which is frequently utilised in classification applications, and the ordinal loss function, which is based on an adjustable penalty matrix. Ordinal loss showed better performances on two different data sets for all architectures.

- **Question 3** What is the effect of the two-stage fine-tuning in KL based knee OA severity classification when we have small size labeled data?

Firstly, we fine-tuned a model to classify the stages of knee OA using the larger size of knee X-ray data collected from multiple centres. To use the features extracted by

the model on this data, we fine-tuned the model again, this time using the smaller size of labelled knee X-ray data collected from a single centre. In **Section 4.4.3**, it was shown that this two-stage fine-tuning process performed better on the smaller size of labelled data.

- **Question 4** How can we leverage unlabeled images to train a deep model to have a robust model?

In **Chapter 5**, we obtained an encoder with a pre-training phase based on contrastive learning. We then integrated this encoder into a supervised fine-tuning phase with a limited number of labelled data. With this technique we show that the proposed pipeline is more robust to a gradual reduction of the number of labelled data.

6.2 Summary and Contributions

The incidence of knee OA has increased within the past several years. This increase has brought along the need for accurate and early diagnosis of the condition. Traditionally, diagnosis of the disease is commonly made through clinical examination and imaging techniques. However, the subjective nature of radiographic findings often leads to inconsistencies in assessments by radiologists. This, in turn, can negatively impact the accuracy of knee OA grading and treatment decisions. Hence, in this research, we aim to utilize deep learning methods to diagnose even the early stages of knee OA. Thus, we aim to better understand the progression of the disease and develop effective treatment strategies. Achieving this goal largely depends on the performance of the models. However, the major obstacle to presenting a robust deep learning model with high performance is the lack of a large amount of labeled data. The absence of labeled knee X-ray data in the literature led us to search for solutions to improve model performance with the existing limited data. Since the problem of accessing labeled data is a common issue for deep learning models, we customize our solutions in this study according to different data scenarios.

- Firstly, we introduced fully supervised methods using labeled data in Chapter 4. To provide solutions with a relatively sufficient number of labeled datasets in Chapter

4.4.1, we presented the test results of SOTA models on the OAI dataset, which is publicly accessible, alongside models commonly used in the literature. Thus, our aim is to demonstrate which model can better extract diagnostic features such as joint space narrowing from knee X-ray data, the presence of osteophytes, and bone marrow lesions to classify the stages of knee osteoarthritis. ConvNeXt V2-Tiny and ConvNeXt V1-Base have yielded the best results in all scores on the OAI test dataset. Additionally, in the same chapter, we presented the superior performance of ConvNeXt V1-Base among three different models trained using in-house hospital data.

- Also, in Chapter 4.4.2, the comparison between the categorical cross-entropy loss function and the ordinal loss revealed that ordinal loss better handles the ordered relationship between KL grades representing disease stages. This is particularly useful in classifying the severity of OA. This is because ordinal loss applies a penalty considering the difference in distance between the true class and the predicted class. The categorical cross-entropy loss function, on the other hand, is not suitable as a loss function for disease severity classification since it only considers the probability value of the correct label and does not penalize the probability values generated for incorrect labels.
- For a scenario with a limited amount of labeled data gathered from a single center, we employ the two-stage fine-tuning method described in Chapter 4.4.3. The limited training data exhibited inadequate performance in detecting KL grades. In response to this scenario, we demonstrate that re-fine-tuning models previously fine-tuned with publicly available OAI data containing data gathered from different centers, using limited number of data from a different center, yielded promising results. Our aim was to increase heterogeneity by pre-learning significant features from a large labeled dataset.
- In Chapter 5, we introduced a semi-supervised method in a scenario where unlabeled data more than labeled data. We propose integrating an encoder, obtained through a pre-training process based on contrastive learning using unlabeled X-ray data, into the supervised fine-tuning stage with a small scale of labeled

data. Utilizing the encoder, which represents features extracted from unlabeled data, during the supervised fine-tuning stage has resulted in an enhancement in model performance. The findings in Chapter 5.2 have provided comprehensive information regarding the effectiveness of the suggested pipeline.

6.3 Limitations and Future Work

Semi-supervised approaches in image-based disease diagnosis represent an open research area where many researchers are concentrating. While we improved the performance of models in scenarios with small amount of labeled data using self-supervised learning, the model still has the capacity for further improvement. So the methods we present in our study have some limitations. First of all, while presenting our semi-supervised approach, we use only one SimCLR method for contrastive learning in the pre-training phase. Additionally, we used only ConvNeXt V1-Base as the encoder architecture in our experiments. Our initial aim is to validate our technique with experiments involving more variations (such as using different encoder architectures, alternative contrastive learning paradigms, or self-supervision paradigms). Additionally, we didn't use different datasets in the pre-training and supervised classification part of our SSL approach. In our future work, we will validate this pipeline using different datasets. We aim to frame our approach with a continual learning scheme on datasets where manual annotations are limited. When creating positive pairs during contrastive learning, commonly used transformations typically enable the extraction of important features for natural images, whereas this may not be the case for medical images where semantic meaning is confined to a very small area on the image. Therefore, we plan to apply a specific method for generating positive pairs tailored to knee X-ray data during the contrastive learning stage. Thus, we aim to improve the pre-training process by producing more meaningful representations using unlabeled data.

REFERENCES

- [1] **Teoh, Y.X., Lai, K.W., Usman, J., Goh, S.L., Mohafez, H., Hasikin, K., Qian, P., Jiang, Y., Zhang, Y. and Dhanalakshmi, S.** (2022). Discovering Knee Osteoarthritis Imaging Features for Diagnosis and Prognosis: Review of Manual Imaging Grading and Machine Learning Approaches, *Journal of healthcare engineering*, 2022, 4138666, retraction published J Healthc Eng. 2023 Oct 11;2023:9765742.
- [2] **Bijlsma, J.W., Berenbaum, F. and Lafeber, F.P.** (2011). Osteoarthritis: An update with relevance for clinical practice, *The Lancet*, 377(9783), 2115–2126.
- [3] **Prezja, F., Annala, L., Kiiskinen, S., Lahtinen, S. and Ojala, T.** (2023). *Adaptive Variance Thresholding: A Novel Approach to Improve Existing Deep Transfer Vision Models and Advance Automatic Knee-Joint Osteoarthritis Classification*, 2311.05799.
- [4] **Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C. and Krishnan, D.** (2021). *Supervised Contrastive Learning*, 2004.11362.
- [5] **Phung, V.H. and Rhee, E.J.** (2019). A High-Accuracy Model Average Ensemble of Convolutional Neural Networks for Classification of Cloud Image Patches on Small Datasets, *Applied Sciences*, 9(21), 4500, <https://doi.org/10.3390/app9214500>.
- [6] **Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N.** (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, 2010.11929.
- [7] **Tiulpin, A., Thevenot, J., Rahtu, E. et al.** (2018). Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach, *Scientific reports*, 8, 1727, <https://doi.org/10.1038/s41598-018-20132-7>.
- [8] **Wang, Z., Chetouani, A., Jarraya, M., Hans, D. and Jennane, R.** (2023). *Transformer with Selective Shuffled Position Embedding and Key-Patch Exchange Strategy for Early Detection of Knee Osteoarthritis*, 2304.08364.
- [9] **Bengio, Y.** (2009). Learning Deep Architectures for AI, *Foundations and Trends® in Machine Learning*, 2(1), 1–127, <http://dx.doi.org/10.1561/22000000006>.

- [10] **Wu, H. and Shapiro, J.L.** (2006). Does overfitting affect performance in estimation of distribution algorithms, *M. Cattolico, editor, Genetic and Evolutionary Computation Conference, GECCO 2006, Proceedings, Seattle, Washington, USA, July 8-12, 2006*, ACM, pp.433–434, <https://doi.org/10.1145/1143997.1144078>.
- [11] **Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T. and Xie, S.** (2022). A ConvNet for the 2020s, *CoRR*, *abs/2201.03545*, <https://arxiv.org/abs/2201.03545>, 2201.03545.
- [12] **Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S. and Xie, S.** (2023). *ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders*, 2301.00808.
- [13] **Chen, T., Kornblith, S., Norouzi, M. and Hinton, G.** (2020). *A Simple Framework for Contrastive Learning of Visual Representations*, 2002.05709.
- [14] **Zhang, L., Hu, J. and Athanasiou, K.** (2009). The role of tissue engineering in articular cartilage repair and regeneration, *Critical reviews in biomedical engineering*, 37(1-2), 1–57, <https://doi.org/10.1615/critrevbiomedeng.v37.i1-2.10>.
- [15] **Sophia Fox, A.J., Bedi, A. and Rodeo, S.A.** (2009). The basic science of articular cartilage: structure, composition, and function, *Sports health*, 1(6), 461–468, <https://doi.org/10.1177/1941738109350438>.
- [16] **Loeser, R.F.** (2006). Molecular mechanisms of cartilage destruction: mechanics, inflammatory mediators, and aging collide, *Arthritis and rheumatism*, 54(5), 1357–1360, <https://doi.org/10.1002/art.21813>.
- [17] **Heidari, B.** (2011). Knee osteoarthritis prevalence, risk factors, pathogenesis and features: Part I, *Caspian journal of internal medicine*, 2(2), 205–212.
- [18] **Wilkins, A.N. and Phillips, E.M.**, (2008). Chapter 61 - Knee Osteoarthritis, **W.R. Frontera, J.K. Silver and T.D. Rizzo**, editors, *Essentials of Physical Medicine and Rehabilitation*, W.B. Saunders, second edition, pp.345–353, <https://www.sciencedirect.com/science/article/pii/B9781416040071500638>.
- [19] **Maqbool, M., Fekadu, G., Jiang, X., Bekele, F., Tolossa, T., Turi, E., Fetensa, G. and Fanta, K.** (2021). An up to date on clinical prospects and management of osteoarthritis, *Annals of medicine and surgery (2012)*, 72, 103077, <https://doi.org/10.1016/j.amsu.2021.103077>.
- [20] **Nguyen, U.S.D.T., Zhang, Y., Zhu, Y., Niu, J., Zhang, B. and Felson, D.T.** (2011). Increasing prevalence of knee pain and symptomatic knee osteoarthritis: survey and cohort data, *Annals of internal medicine*, 155(11), 725–732.

- [21] **Palo, N., Chandel, S., Dash, S., Arora, G., Kumar, M. and Biswal, M.** (2015). Effects of osteoarthritis on quality of life in elderly population of Bhubaneswar, India: A prospective multicenter screening and therapeutic study of 2854 patients, *Geriatric orthopaedic surgery & rehabilitation*, 6(4), 269–275.
- [22] **Dantas, L.d.O., Salvini, T.F. and McAlindon, T.E.** (2021). Knee osteoarthritis: Key treatments and implications for physical therapy, *Brazilian journal of physical therapy*, 25(2), 135–146, <https://doi.org/10.1016/j.bjpt.2020.08.004>.
- [23] **Lim, W.B. and Al-Dadah, O.** (2022). Conservative treatment of knee osteoarthritis: A review of the literature, *World journal of orthopedics*, 13(3), 212–229, <https://doi.org/10.5312/wjo.v13.i3.212>.
- [24] **Braun, H. and Gold, G.** (2012). Diagnosis of osteoarthritis: imaging, *Bone*, 51(2), 278–288, <https://doi.org/10.1016/j.bone.2011.11.019>.
- [25] **Kohn, M.D., Sassoon, A.A. and Fernando, N.D.** (2016). Classifications in Brief: Kellgren-Lawrence Classification of Osteoarthritis, *Clinical orthopaedics and related research*, 474(8), 1886–1893.
- [26] **Kellgren, J.H. and Lawrence, J.S.** (1957). Radiological assessment of osteo-arthritis, *Annals of the rheumatic diseases*, 16(4), 494–502, <https://doi.org/10.1136/ard.16.4.494>.
- [27] **Panetta, D.**, (2014). Computed Tomography, **A. Brahme**, editor, *Comprehensive Biomedical Physics*, Elsevier, pp.65–88.
- [28] **Piccolo, C.L., Mallio, C.A., Vaccarino, F., Grasso, R.F. and Zobel, B.B.** (2023). Imaging of knee osteoarthritis: a review of multimodal diagnostic approach, *Quantitative imaging in medicine and surgery*, 13(11), 7582–7595, <https://doi.org/10.21037/qims-22-1392>.
- [29] **Koff, D.A. and Doyle, T.E.**, (2019). Imaging Informatics, **R. Narayan**, editor, *Encyclopedia of Biomedical Engineering*, Elsevier.
- [30] **Khairunnahar, L., Hasib, M.A., Rezanur, R.H.B., Islam, M.R. and Hosain, M.K.** (2019). Classification of malignant and benign tissue with logistic regression, *Informatics in Medicine Unlocked*, 16, 100189, <https://doi.org/10.1016/j.imu.2019.100189>.
- [31] **S, R., Bharadwaj, A.S., K, D.S., Khadabadi, M.S. and Jayaprakash, A.** (2022). Digital Implementation of the Softmax Activation Function and the Inverse Softmax Function, *2022 4th International Conference on Circuits, Control, Communication and Computing (I4C)*, pp.64–67.
- [32] **Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M. and Lempitsky, V.** (2016). *Domain-Adversarial Training of Neural Networks*, 1505.07818.

- [33] **Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H. and Tao, D.** (2023). *A Survey on Self-supervised Learning: Algorithms, Applications, and Future Trends*, 2301.05712.
- [34] **Olaoye, F., Potter, K. and Doris, L.** (2024). Computer Vision and Image Recognition Techniques, *Journal of Scientific Conference Proceedings*.
- [35] **O'Shea, K. and Nash, R.** (2015). *An Introduction to Convolutional Neural Networks*, 1511.08458.
- [36] **Castellino, R.A.** (2005). Computer aided detection (CAD): an overview, *Cancer imaging : the official publication of the International Cancer Imaging Society*, 5(1), 17–19.
- [37] **Cicerone, M.T. and Camp Jr., C.H.**, (2019). Potential Roles for Spectroscopic Coherent Raman Imaging for Histopathology and Biomedicine, *Neurophotonics and Biomedical Spectroscopy*, Elsevier, pp.547–570.
- [38] (2020). *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1*, Springer International Publishing, <http://dx.doi.org/10.1007/978-3-030-17795-9>.
- [39] **Guo, Z., Xie, J., Wan, Y., Zhang, M., Qiao, L., Yu, J., Chen, S., Li, B. and Yao, Y.** (2022). A review of the current state of the computer-aided diagnosis (CAD) systems for breast cancer diagnosis, *Open life sciences*, 17(1), 1600–1611.
- [40] **Qin, C., Yao, D., Shi, Y. and Song, Z.** (2018). Computer-aided detection in chest radiography based on artificial intelligence: a survey, *Biomedical engineering online*, 17(1), 113, <https://doi.org/10.1186/s12938-018-0544-y>.
- [41] **Jamshidi, A., Pelletier, J.P. and Martel-Pelletier, J.** (2019). Machine-learning-based patient-specific prediction models for knee osteoarthritis, *Nature reviews. Rheumatology*, 15(1), 49–60, <https://doi.org/10.1038/s41584-018-0130-5>.
- [42] **Kluzek, S. and Mattei, T.A.** (2019). Machine-learning for osteoarthritis research, *Osteoarthritis and cartilage*, 27(7), 977–978, <https://doi.org/10.1016/j.joca.2019.04.005>.
- [43] **Nelson, A.E., Fang, F., Arbeeve, L., Cleveland, R.J., Schwartz, T.A., Callahan, L.F., Marron, J. and Loeser, R.F.** (2019). A machine learning approach to knee osteoarthritis phenotyping: data from the FNIH Biomarkers Consortium, *Osteoarthritis and cartilage*, 27(7), 994–1001, <https://doi.org/10.1016/j.joca.2018.12.027>.
- [44] **Brahim, A., Jennane, R., Riad, R., Janvier, T., Khedher, L., Toumi, H. and Lespessailles, E.** (2019). A decision support tool for early detection of knee OsteoArthritis using X-ray imaging and machine learning: Data from

the OsteoArthritis Initiative, *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 73, 11–18.

- [45] **Kotti, M., Duffell, L.D., Faisal, A.A. and McGregor, A.H.** (2017). Detecting knee osteoarthritis and its discriminating parameters using random forests, *Medical engineering & physics*, 43, 19–29, <https://doi.org/10.1016/j.medengphy.2017.02.004>.
- [46] **Antony, J., McGuinness, K., Connor, N.E.O. and Moran, K.** (2016). *Quantifying Radiographic Knee Osteoarthritis Severity using Deep Convolutional Neural Networks*, 1609.02469.
- [47] **Leung, K., Zhang, B., Tan, J., Shen, Y., Geras, K.J., Babb, J.S., Cho, K., Chang, G. and Deniz, C.M.** (2020). Prediction of Total Knee Replacement and Diagnosis of Osteoarthritis by Using Deep Learning on Knee Radiographs: Data from the Osteoarthritis Initiative, *Radiology*, 296(3), 584–593, <https://doi.org/10.1148/radiol.2020192091>.
- [48] **Chen, P., Gao, L., Shi, X., Allen, K. and Yang, L.** (2019). Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss, *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 75, 84–92, <https://doi.org/10.1016/j.compmedimag.2019.06.002>.
- [49] **Nasser, Y., Jennane, R., Chetouani, A., Lespessailles, E. and Hassouni, M.E.** (2020). Discriminative Regularized Auto-Encoder for Early Detection of Knee OsteoArthritis: Data from the Osteoarthritis Initiative, *IEEE transactions on medical imaging*, 39(9), 2976–2984.
- [50] **Maurício, José & Domingues, I.B.J.** (2023). Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review, *Applied Sciences*, 13, 5521.
- [51] **Nguyen, H.H., Saarakkala, S., Blaschko, M.B. and Tiulpin, A.** (2022). CLIMAT: Clinically-Inspired Multi-Agent Transformers for Knee Osteoarthritis Trajectory Forecasting, *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pp.1–5.
- [52] **Berrimi, M., Hans, D. and Jennane, R.** (2024). A semi-supervised multiview-MRI network for the detection of Knee Osteoarthritis, *Computerized Medical Imaging and Graphics*, 114, 102371, <https://www.sciencedirect.com/science/article/pii/S089561112400048X>.
- [53] **Alam, M., Yamashita, R., Ramesh, V. et al.** (2023). Contrastive learning-based pretraining improves representation and transferability of diabetic retinopathy classification models, *Scientific Reports*, 13(6047), <https://doi.org/10.1038/s41598-023-33365-y>.

- [54] **Hossain, M.B., Iqbal, S.M.H.S., Islam, M.M., Akhtar, M.N. and Sarker, I.H.** (2022). Transfer learning with fine-tuned deep CNN ResNet50 model for classifying COVID-19 from chest X-ray images, *Informatics in Medicine Unlocked*, 30, 100916.
- [55] **Wang, Y., Gu, X., Hou, W., Zhao, M., Sun, L. and Guo, C.** (2023). Dual Semi-Supervised Learning for Classification of Alzheimer’s Disease and Mild Cognitive Impairment Based on Neuropsychological Data, *Brain Sciences*, 13(2), 306.
- [56] **Calivá, F., Kamat, S., Morales Martinez, A., Majumdar, S. and Pedoia, V.** (2022). Surface Spherical Encoding and Contrastive Learning for Virtual Bone Shape Aging, *Medical Image Analysis*, 77, 102388.
- [57] **He, K., Zhang, X., Ren, S. and Sun, J.** (2016). Deep Residual Learning for Image Recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.770–778.
- [58] **Xie, S., Girshick, R.B., Dollár, P., Tu, Z. and He, K.** (2016). Aggregated Residual Transformations for Deep Neural Networks, *CoRR*, *abs/1611.05431*, <http://arxiv.org/abs/1611.05431>, 1611.05431.
- [59] **Huang, G., Liu, Z., van der Maaten, L. and Weinberger, K.Q.** (2018). *Densely Connected Convolutional Networks*, 1608.06993.
- [60] **Wang, S. and Zhang, Y.** (2020). DenseNet-201-Based Deep Neural Network with Composite Learning Factor and Precomputation for Multiple Sclerosis Classification, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16, 1 – 19, <https://api.semanticscholar.org/CorpusID:225557866>.
- [61] **Tan, M. and Le, Q.V.** (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, *CoRR*, *abs/1905.11946*, <http://arxiv.org/abs/1905.11946>, 1905.11946.
- [62] **Niu, Z., Zhou, M., Wang, L., Gao, X. and Hua, G.** (2016). Ordinal Regression with Multiple Output CNN for Age Estimation, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.4920–4928.
- [63] **Ferrer, L.** (2023). *Analysis and Comparison of Classification Metrics*, 2209.05355.
- [64] **Maier-Hein, L., Reinke, A., Godau, P. and Tizabi, M.D.** (2023). *Metrics reloaded: Recommendations for image analysis validation*, 2206.01653.
- [65] **Galdran, A.** (2023). *Performance Metrics for Probabilistic Ordinal Classifiers*, 2309.08701.
- [66] **Kingma, D.P. and Ba, J.** (2017). Adam: A Method for Stochastic Optimization, 1412.6980.

- [67] **Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J. and Han, J.** (2021). *On the Variance of the Adaptive Learning Rate and Beyond*, 1908.03265.
- [68] **Caruana, R., Lawrence, S. and Giles, L.** (2001). Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping, *Advances in neural information processing systems*, pp.402–408.
- [69] **Nevitt, M., Felson, D. and Lester, G.** (2006). The Osteoarthritis Initiative: Protocol for the Cohort Study, *Proceedings of the Osteoarthritis Initiative Study*.
- [70] **Chen, P.** (2018). *Knee Osteoarthritis Severity Grading Dataset*, <https://data.mendeley.com/datasets/56rmx5bjcr/1>.
- [71] **Deng, J., Dong, W., Socher, R., Li, L.J. and Li, L.F.F.** (2009). ImageNet: A large-scale hierarchical image database, *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, pp.248–255.
- [72] **Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D.** (2019). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, *International Journal of Computer Vision*, 128(2), 336–359, <http://dx.doi.org/10.1007/s11263-019-01228-7>.



CURRICULUM VITAE

Name Surname: İlknur AKTEMUR

EDUCATION:

- **B.Sc.:** 2020, Istanbul Kultur University, Engineering Faculty, Computer Engineering Department

PROFESSIONAL EXPERIENCE AND REWARDS:

- ML Engineer at Novus Writer (2023-...)
- ML and AI Engineer at FIGES Engineering (2021-2023)

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- **Aktemur I., Oksuz I.** Automatic Detection of Knee Osteoarthritis Severity with SOTA Deep Learning Models and Ordinal Loss (2023)- 31st Signal Processing and Communications Applications Conference (SIU), Istanbul, Turkey.

OTHER PUBLICATIONS, PRESENTATIONS AND PATENTS:

- **Aktemur I.**, Erensoy K. and Kocyigit E., Optimization of Waste Collection in Smart Cities with the use of Evolutionary Algorithms (2020)- International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey

