KNEE OSTEOARTHRITIS GRADING USING DEEP LEARNING CLASSIFIER

By

Jevyline Ng

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF COMPUTER SCIENCE (HONOURS)
Faculty of Information and Communication Technology
(Kampar Campus)

JUNE 2024

UNIVERSITI TUNKU ABDUL RAHMAN

REPORT STATUS DECLARATION FORM

Title: KNEE OSTEOARTHRITIS GRADING USING DEEP LEARNING CLASSIFIER

Academic Session: JUN 2024

I JEVYLINE NG

(CAPITAL LETTER)

declare that I allow this Final Year Project Report to be kept in

Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

1. The dissertation is a property of the Library.
2. The Library is allowed to make copies of this dissertation for academic purposes.

Verified by,

[illegible]	[illegible]
(Author's signature)	(Supervisor's signature)
Address:	
95, Hala Menglembu Timur 2,	
31450 Ipoh,	
Perak	Ts. Dr. Goh Chuan Meng
	Supervisor's name
Date: 13/09/2024	**Date**: 13/09/2024

Universiti Tunku Abdul Rahman			
Form Title : **Sample of Submission Sheet for FYP/Dissertation/Thesis**			
Form Number: **FM-IAD-004**	Rev No.: **0**	Effective Date: **21 JUNE 2011**	Page No.: **1 of 1**

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

UNIVERSITI TUNKU ABDUL RAHMAN

Date: 13/09/2024

SUBMISSION OF FINAL YEAR PROJECT /DISSERTATION/THESIS

It is hereby certified that **JEVYLINE NG** (ID No: **2100637**) has completed this final year project entitled "**KNEE OSTEOARTHRITIS GRADING USING DEEP LEARNING CLASSIFIER**" under the supervision of **DR. TS. GOH CHUAN MENG** (Supervisor) from the Department of **COMPUTER SCIENCE**, Faculty of **INFORMATION AND COMMUNICATION TECHNOLOGY**.

I understand that University will upload softcopy of my final year project in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.

Yours truly,

[illegible]

(JEVYLINE NG)

*Delete whichever not applicable

DECLARATION OF ORIGINALITY

I declare that this report entitled "**METHODOLOGY, CONCEPT AND DESIGN OF A 2-MICRON CMOS DIGITAL BASED TEACHING CHIP USING FULL-CUSTOM DESIGN STYLE**" is my own work except as cited in the references. The report has not been accepted for any degree and is not being submitted concurrently in candidature for any degree or other award.

Signature	:	[illegible]
Name	:	JEVYLINE NG
Date	:	13/09/2024

ACKNOWLEDGEMENTS

I would like to express my sincere thanks and appreciation to my supervisor, Dr. Ts. Goh Chuan Meng who has given me this bright opportunity to engage in a Deep Learning Model research project. It is my first step to establish a career in Deep Learning. A million thanks to you.

I would also like to express sincere appreciation to the medical professionals and institution that generously provided access to the X-ray images dataset, without which this study would not have been possible.

Lastly, I would like to thank my family and friends for their understanding and patience during the course of this research. Their encouragement and support have been a constant source of motivation.

ABSTRACT

This project is a Deep Learning Model research project for academic purpose. Knee osteoarthritis (OA) is a prevalent degenerative joint disorder affecting millions worldwide, leading to pain, impaired mobility, and diminished quality of life. Accurate grading of OA severity is crucial for effective clinical management, yet it remains a challenging task prone to subjectivity and inter-observer variability. In this study, we propose a novel approach utilizing deep learning classifiers to automate the grading process of knee OA based on the Kellgren Lawrence grading system. Through the development and evaluation of multiple deep learning models, we aim to provide a robust and reliable tool for clinicians to objectively assess OA severity from X-ray images. Our methodology involves preprocessing of X-ray images, followed by feature extraction and classification using convolutional neural networks (CNNs). The performance of each model is assessed through rigorous validation on a diverse dataset of knee X-ray images annotated with ground truth Kellgren Lawrence grades.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

OA	Osteoarthritis
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
CNN	Convolutional Neural Network
DLC	Deep Learning Classifier
ROI	Region of Interest
DenseNet	Densely Connected Convolutional Networks
VGGNet	Visual Geometry Group Network
ResNet	Residual Network
GoogLeNet	Google's Inception Network
RegNet	Regularized Convolutional Neural Network
GradCAM	Gradient-weighted Class Activation Mapping
ROC	Receiver Operating Characteristic
AUC	Area Under Curve
NCA	Neighborhood Components Analysis

Chapter 1 Introduction

1.1 Motivation and Problem Statement

Osteoarthritis (OA) is one of the most common joint diseases, particularly lower extremity joints such as knee and hip, eventually leading to disability in the lower limbs [1]. Most people thought that OA will only affect the aging community but in fact it affects most of the population. Study says that genetics, obesity, diet, physical activity, knee injuries are also potential risks for the OA incidence [1]. Even though the disease is still considered incurable in the medical field, there are chances to mitigate the deterioration of the disease if detected in the early stage [2].

Since the first radiography, X-ray introduced in the late 19th century by Wilhelm Conrad Roentgen, it has been widespread and developed throughout the 100 years resulting in more medical imaging techniques such as Fluoroscopy, Computed Tomography (CT) Scanning and Magnetic Resonance Imaging (MRI) [3]. The advancement in medical imaging have enhanced ability of healthcare professionals to perform diagnostics and formulate optimal treatment plan for patient. However, complexity of the knee joint structure and diversity of each human body caused manual detection by healthcare professionals consume more time. There have been active studies regarding detection of the disease throughout the years and implementation of Artificial Intelligence in medical field since recent years is the most potential way in successful early detection. Even so, further research in the field is imperative to refine and optimize these AI-based detection model to ensure the reliability and applicability across diverse patient population. Full automation of OA detection will be a remarkable achievement in both AI and medical field as it can speed up the diagnostics process and lower cost of the disease detection making the healthcare resource more accessible and sustainable.

1.2 Research Objectives

i. **Find the best deep learning classifier to grade severity of OA on the knee X-ray images.**

   Run a comparative analysis of several deep learning models to evaluate the performance of various automated grading methods for knee OA based on the KL grading system.

ii. **Optimize the deep learning model to enhance its accuracy and sensitivity in identifying OA features on X-ray images.**

   Perform various optimization strategies to achieve a better model including architecture selection, hyperparameter tuning, data augmentation and more.

iii. **Implementing transfer learning to leverage the model overall performance**

   Transfer learning allows model to converge faster during training and benefits in term of improving model generalization ability, reducing data requirements and having better feature representation.

1.3 Project Scope and Direction

The project will involve a systematic approach to develop an automated system for knee OA diagnosis and severity classification based on the KL grading using deep learning techniques.

Firstly, a diverse dataset of knee X-ray images labelled with the OA severity grades will be used. The dataset will serve as the foundation for training and evaluating the deep learning model. The images will undergo preprocessing steps including standardization, resizing and normalization to enhance the model performance and ensure consistency across the dataset.

Next, various deep learning architectures, particularly Convolutional Neural Networks (CNNs), will be experimented with to develop an accurate OA classifier. The model will be trained using labelled data and validated using cross-validation techniques to ensure robustness and generalizability. A comparative analysis will be conducted among various models to identify the one that demonstrates superior performance compared to other models.

Following model development, optimization techniques will be employed to fine-tune hyperparameters and explore transfer learning methods to enhance the model's performance. This includes adjusting learning rates, batch sizes and exploring pre-trained models to leverage knowledge from related tasks.

For evaluation, quantitative metrics such as accuracy, sensitivity and specificity will be assessed to quantify the model's performance. The automated classification results will be compared with manual grading by health professionals to evaluate the efficacy of the model.

Finally, the project will deliver a trained deep learning model capable of diagnosing and classifying severity of OA based on the patient's knee X-ray images. Accompanying documentation detailing the model architecture, training process and evaluation results will be provided.

1.4 Impact, Significance and Contribution

The implementation of deep learning models for knee OA diagnosis and severity classification holds significant implications for the healthcare industry. By leveraging advanced AI techniques such as CNN and transfer learning, the project aims to revolutionize the current diagnostic process. With the ability to automatically analyse knee X-ray images and accurately classify OA severity, the model could streamline clinical workflows, reduce diagnostic errors and enhance patient outcomes. This helps to alleviate the burden on healthcare professionals, improve treatment decision making and contribute to more effective and efficient management of knee OA.

Furthermore, the research contributes to the growing field of medical imaging and AI-assisted diagnostics. By exploring the efficacy of deep learning algorithms in the context of knee OA diagnosis, the project expands the knowledge base and paves the way for future advancements in automated disease detection systems. The insights gained from this study could inform the development of similar models for other musculoskeletal disorders and facilitate the adoption of AI-driven solutions in clinical practice.

1.5 Report Organization

The details of this research are in the following chapters. In chapter 2, related papers according to deep learning models to grade OA severity are reviewed thoroughly and how the deep learning model works will be discussed. Chapter 3 will be about the system requirements and the system design. Chapter 4 will include the visualization of the work done. Chapter 5 will include the final model's training and testing results. Chapter 6 will conclude the project progress done so far and what can be expected in the future.

Chapter 2 Literature Review

2.1 Literature Review

2.1.1 Classification using CNN

Schwartz et al. [4] researched on classifying OA using a classic CNN model to inspect if the AI-based classifier can perform as well as fellowship-trained knee arthroplasty surgeons. The researchers used combination of CNN training and statistical analysis to develop the model. The model performance is compared with manual grading by 4 expertise in the field. The output of the model shows little deviation with output of manual grading by the 4 experts even though the most basic model is used. This indicated possibilities of AI-based classifier to outperform human in the task with reduced time-consumption.

2.1.2 Classification using Deep Siamese CNN

Tiulpin et al. [5] developed a model employing deep Siamese CNN to predict the severity of OA. The model emphasizes on disease relevant features commonly used in clinical practice such as bone shape and joint space. Traditional Siamese Network takes in 2 image input and measures similarity or dissimilarity, it consists of two identical subnetworks that share the same weight and architecture for each input. The author uses same concept but instead of taking 2 image of input, image symmetry is utilized, and the input taken is equal left and right segment of an image. The author highlighted that utilization of deep Siamese neural network that uses a smaller number of training parameters have improved the robustness of model which can be testament by high AUC score of 0.93. Although the state-of-art approach seems to be promising, the high reliability on large input data makes it prone to overfitting when applied to limited amount of data.

Nguyen et al. [6] further improve the architecture of Tiulpin et al. [5] by implementing semi supervised learning, Semixup on top of the deep Siamese CNN to address the use of huge dataset in the previous research. It also addresses another issue in the previous study regarding potential loss of fine-grained information for knee OA grading. Semixup achieves comparable performance to a well-tuned fully supervised baseline with a balanced accuracy of 71% ± 0.8% despite requiring only one sixth of the labelled data. However, effectiveness of the study may be strongly reliable on the quality of dataset.

2.1.3 Classification using DenseNet

Norman et al. [7] uses a U-Net model to localize left and right knee joints within the radiographs to isolate the ROI and then uses DenseNet neural network architecture to train the model in predicting the severity of OA. Unlike classic CNNs that uses forward pass approach, DenseNet architecture appear to be more complex where dense connections are established between all layers to reduce risk of information loss. The model shows sensitivity rate of 83.7% for no OA, 70.2% for mild OA, 68.9% for moderate OA and 86.0% for severe OA with specificity rates of 86.1%, 83.8%,97.1% and 99.1% respectively. The existence of deviation between sensitivity rate for each level shows the limitation of the model especially when the aim for research is for early detection which ranges in the mild and moderate OA.

Pedoia et al. [8] also proposed the similar deep learning approach, DenseNet that can distinguish radiography with and without OA. On top of the basic DenseNet, the model also studies a feature of MRI which is raw T2 data that refers to the T2 relaxation time of tissues in human body. The study found that DenseNet easily attain AUC of 83.44% when trained on the T2 data compared to conventional shallow model that can only achieve AUC of 77.77%. However, T2 data can only be obtained from MRI that is not as accessible as X-ray.

S. Aslan [9] proposed method of integrating CNN for deep feature extraction, NCA for feature selection and ML models for classification task to grade the knee X-ray images. The CNN architectures are mainly used for feature extraction which will be flowed into NCA for feature selection and lastly using machine learning classifiers to complete the classification task. Among all experimented combinations, the DenseNet201 incorporated with SVM obtained the highest accuracy of 79.3% while other models also obtained nearly the same results, averagely around 77%. However, the achievement does not purely come from the power of the CNN architectures but incorporated with NCA feature selector and variety of machine learning models to accomplish the classification tasks.

2.1.4 Classification ResNet variants

Olsson et al. [10] uses a supervised learning approach with ResNet neural network having 35 layers and batch normalization for each convolutional layer. The chosen ResNet variant has a strong performance due to its ability to handle very deep network architecture while maintaining ease of training. The input image is only pre-processed to fit image to 256 x 256 pixels and not cleaning it from major visual disturbance. The author incorporated 5% of white noise in model training making the model to handle variations and imperfections in real-world data better. The model yielded overall high AUC of more than 0.8. The author stated the research have limitation regarding inclusiveness of different DLC architecture as only ResNet is used in the research.

Kim et al. [11] applied the SE-ResNet architecture to perform grading of OA severity with 2 stages of model training. Firstly, a stack of six squeeze-and-excitation ResNet module was constructed to form a CNN which is used to train images only and predict the OA grade from 0 to 4. Then, the output is taken as input in the next step together with the clinical information into a neural network to improve the accuracy. The study achieved quite high AUC score which are 0,97, 0.85, 0.75, 0.86 and 0.95 for OA grading from 0 to 4 respectively.

However, the clinical data does not cover more possibilities such as past injuries and so on. Even with more related clinical data incorporated, it will only increase burden of model training as such dataset is hard to obtain.

Tiulpin & Saarakkala [12] uses SE-ResNet-50 with ResNetXt blocks as their network architecture which involves an ensemble of residual network with 50 layers. Unlike others, the study uses 2 types of OA grading system which are Kellgren-Lawrence grading and OA Research Society International grading. The method yielded an amusing result which has area under the ROC curve of 0.98 and average precision of 0.98 for detecting presence of OA. The author stated that they faced limitation where bias in performance of the algorithm is possible as the model does not perform as well on the test set compared to the whole dataset.

2.1.5 Classification using VGGNet variants

Górriz et al. [13] proposes an end-to-end neural network architecture which uses a pre-trained VGG-16 network as a base model. The VGG-16 that is popular for its simplicity, consists of 13 convolutional layers and 3 fully connected layers which uses small 3x3 convolutional filters. The author adds a multi-loss training on top of the base model to manage the training of multiple attention branches with differing convergence rates. This approach enhances overall performance by combining attention features with varying level of abstraction. However, the model does not achieve high accuracy with test accuracy of 64.3%.

Anthony et al. [14] implemented 2 variants of VGGNet in their study in classifying severity of OA which are VGG16 and VGG-M-128, a customized version of VGGNet. The model starts with a pre-trained model on ImageNet and repurpose it the to the current classifying task. After that, a linear SVM is employed for classifying the severity of knee OA. Unlike others, the paper suggested that the severity of OA should be treated as a continuous variable instead of the normal multi-class classification. The VGG-M-128 outperform the

VGG16 network architecture and achieved even higher accuracy after fine-tuned. Due to the reliance on SVM for final output, the author doesn’t fully utilize CNNs but only improve over the template matching technique.

Chen et al. [15] evaluated total of 13 popular CNN classifier with variety of network architectures such as ResNet, VGG and DenseNet et al. Gradient-weighted Class Activation Mapping (GradCAM) technique is applied to the model to help identify crucial region for the OA grading. Four ordinal matrices are compared to define the penalty for misclassification between neighbouring grades and it successfully improve accuracy by reducing the mean absolute error (MAE). They found that fine-tuned VGG-19 model with the proposed ordinal loss obtain the best classification accuracy of 69.7% and MAE of 0.344. The VGG-19 network consist of 16 convolutional layers, 3 fully connected layers with 3x3 convolutional filters and 2x2 max-pooling layers. However, a potential limitation of this approach lies in the reliance on manually defined ordinal metrics which might not capture the full complexity of the underlying data making the model lack generalization ability.

2.2 Summary of Literature Review

Table 2.1 Summary of Literature Review

Author	**Medical Imaging Technique**	**Network Architecture**	**Pros**	**Cons**
Schwartz et al. [4]	X-ray	CNN	Consume less processing time to classify same amount of knee images compared to human using basic CNN	Does not include more neural network architecture. Removed unclear images from dataset, lack of generalization in real-world application
Tiulpin et al. [5]	X-ray	Deep Siamese CNN	Taking left and right symmetry of image as model input, reduced training parameter increase efficiency	High reliability on large input data makes it prone to overfitting when applied to limited amount of data
Nguyen et al. [6]	X-ray	Deep Siamese CNN	Implement semi supervised learning that only used 1/6 of labelled training data compared to previous studies and achieved comparable performance	Effectiveness of the study may be strongly reliable on the quality of dataset
Norman et al. [7]	X-ray	DenseNet	Reduce risk of information loss as each layer are densely connected	Existence of deviation between sensitivity rate for each grade especially in mild and moderate may cause early detection to fail

Pedoia et al. [8]	MRI	DenseNet	Usage of extra clinical information, T2 provided more information to the model	Rely on T2 data that is only provided in MRI, hard to obtain dataset and is less common compared to X-ray
Serpil Aslan [9]	X-ray	DenseNet201	Use deep learning model for feature extraction, NCA for feature selection and SVM for final classification task.	
Olsson et al. [10]	X-ray	ResNet	Uses minimum image preprocessing steps and implement white noise to make model have strong generalization ability	Lack inclusiveness in DLC as only one model is used
Kim et al. [11]	X-ray	SE-ResNet	Use of more clinical data helps the model to learn from more aspects	Rely on extra clinical data aside from images, hard to obtain training dataset and too much information increases computational power
Tiulpin & Saarakkala [12]	X-ray	SE-ResNet-50	Train the model with 2 types of OA grading system makes the model applicable in both grading system	Bias in performance of the algorithm is possible as the model does not perform as well on the test set compared to the whole dataset
Górriz et al. [13]	X-ray	VGG-16	Simplicity of base model, added multi-loss training on top of the base model to	Does not achieve high accuracy

			manage the training of multiple attention branches	
Anthony et al. [14]	X-ray	VGG-16 &VGG-M-128	Creative approach where severity of OA is treated as a continuous variable, comparing multiple DLC to obtain the best	Rely on SVM for final output, does not fully utilize neural network architecture
Chen et al. [15]	X-ray	VGG-19	Implement GradCAM to identify crucial region, high inclusiveness as 13 DLC models are compared to obtain best performance	Manually defined ordinal metrics might not capture the full complexity of the underlying data making the model lack generalization ability

After going through different types of DLC and reviewing all these papers, there are few things that must be taken note of in the research and model development phase. Some of the major things are:

- Dataset selection is important, ensuring accurate annotations and representative samples can enhance model performance.
- The need of model architecture exploration, experimenting with various deep learning architectures allows for identifying the most suitable architecture for the specific task.
- Having standard preprocessing techniques is important, implementing standard preprocessing techniques improve the quality of input data and speed up model training process.
- The need of exploring advanced preprocessing techniques that can leverage the input data quality, especially the preprocessing method tailored to medical imaging.
- Most of the models used in medical imaging implements transfer learning, incorporating transfer learning strategies from related medical imaging enhance the robustness of model.
- Setting up for suitable evaluation metrics, appropriate evaluation metrics such as accuracy, sensitivity and specificity should be used to assess the model performance.

Some non-major things to take note of:

- Try to use data augmentation techniques if having problem of dataset that is too small, this can help to enhance generalization ability of model.
- Ensemble learning can be implemented by combining predictions from multiple models.

By carefully considering these critical features during the research and model development phase, we can significantly enhance the efficiency and effectiveness of our deep learning classifier.

Chapter 3 System Model

3.1 System Requirement

3.1.1 Software

Table 3.1 Software Involved

Software	Version	Function
Python	**3.10.14**	Serves as the primary programming language used for coding all of the image preprocessing and model development process. It offers extensive libraries and frameworks for data manipulation and scientific computing.
Pytorch	**2.4.0+cpu**	Open-source deep learning framework. Mainly used in implementing and training all the models in the research. Source of all pretrained model used.
Opencv	**4.8.0**	Provides powerful tools for image processing enabling preprocessing and feature extraction from input images. Used for image sharpening in the research.
Kaggle	**1.6.17**	Software used to do all the coding. Source of dataset used. Provide GPU that is a must for training all the heavy models.

3.1.2 Hardware

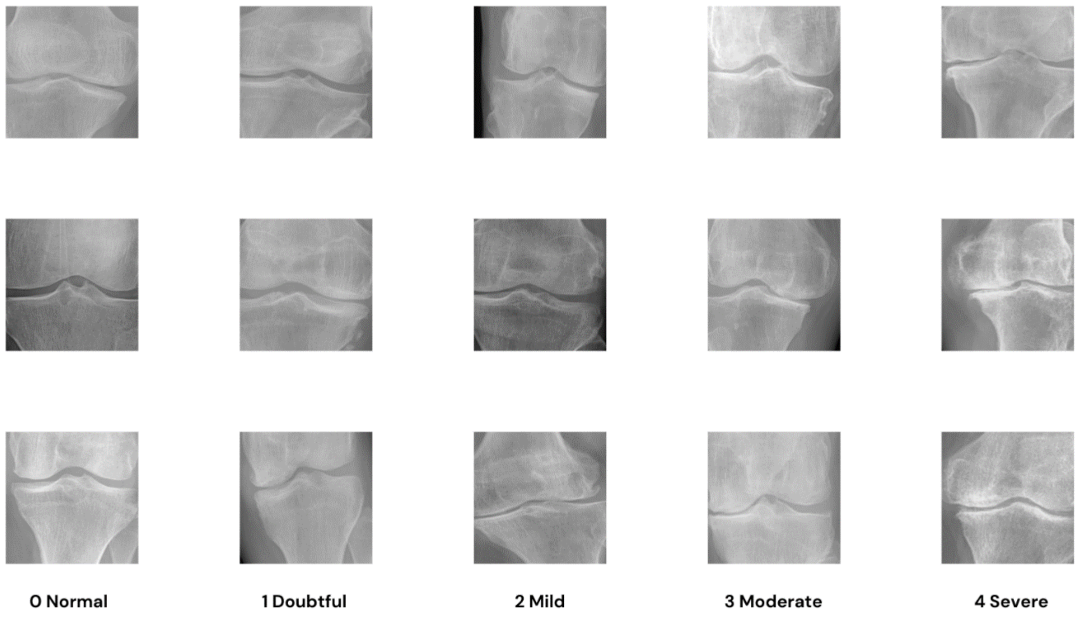
Table 3.2 Laptop Specifications

Description	Specifications
Model	Huawei MateBook D15
OS	Windows 11 Home Single Language
CPU	AMD Ryzen 7 3700U with Radeon Vega Mobile Gfx @ 2.30 GHz
RAM	8.00 GB

3.2 System Overview

3.2.1 Dataset

The dataset used in this research was obtained from open source Kaggle - “Digital Knee X-ray” [16] which is similar to the dataset used by S. Aslan [9] in the research, as the OAI (Osteoarthritis Initiative) and MOST (Multicenter Osteoarthritis Study) datasets were not directly accessible for public use. The dataset consists of 1,650 knee X-ray images, each graded by two independent medical experts for osteoarthritis (OA) severity. For the purpose of this study, only the grading provided by Medical Expert I was utilized. Since the images was all in 1 folder, train test validation split with ratio train (70%), validation (15%) and test (15%) were done. Each image was categorized into one of five classes based on the severity of OA: 0



(Normal), 1 (Doubtful), 2 (Mild), 3 (Moderate), and 4 (Severe).

Figure 3.2.1 Sample X-ray images from dataset

3.2.2 Model Development Flowchart

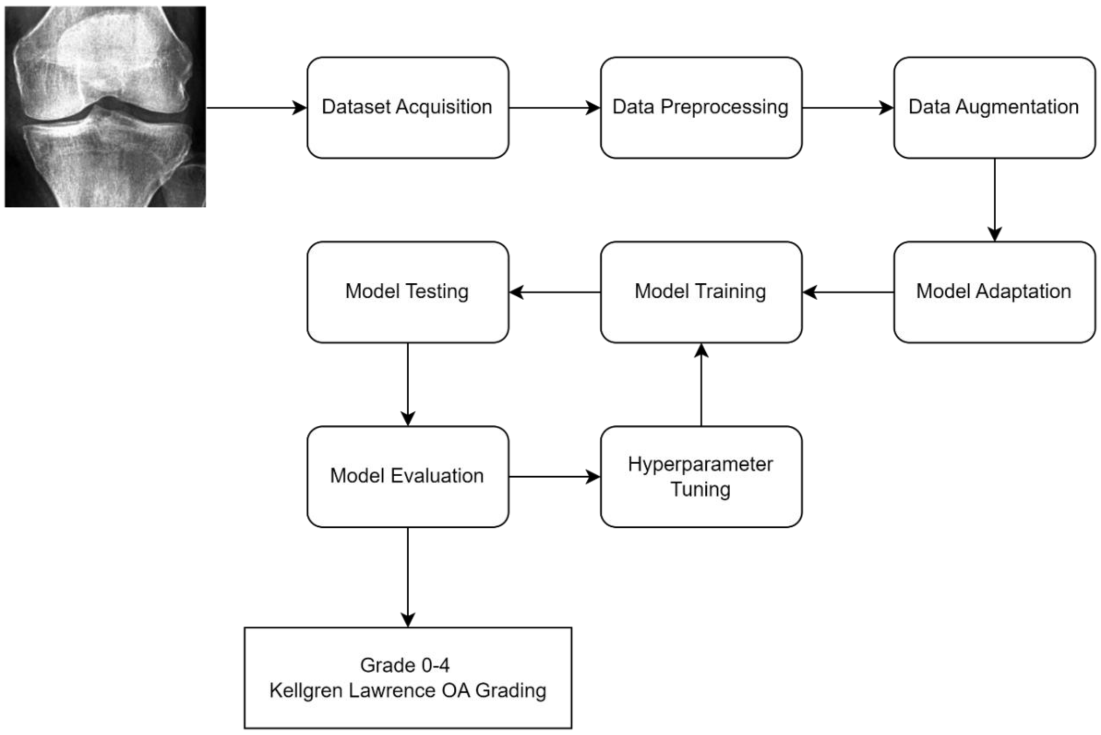


Figure 3.2.2 Flowchart for model development

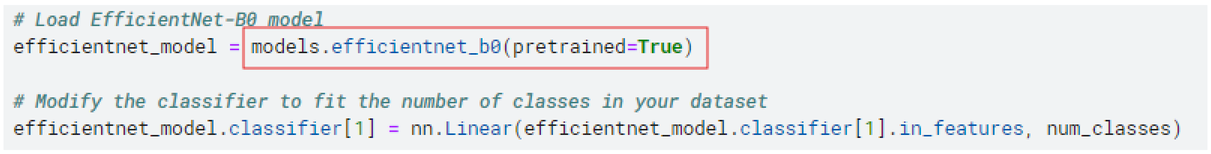
The model development flowchart outlines a systematic approach to developing and optimizing machine learning models. It begins with dataset acquisition, where relevant data is collected for the task at hand. This step is crucial as the quality and quantity of data directly impact model performance. Following dataset acquisition, data preprocessing is performed to clean, normalize, and format the data to make it suitable for training. This step involves handling missing values, scaling features, and encoding categorical variables to prepare the data for model training.

After data preprocessing, data augmentation techniques may be applied to increase the diversity of the dataset and improve model generalization. This involves generating synthetic data by applying transformations such as rotation, flipping, or scaling to the existing data samples. Once the dataset is prepared, the model adaptation phase involves selecting an appropriate architecture and configuring it to suit the specific problem domain. This includes defining the layers, activation functions, and optimization algorithms for the model.

With the model architecture defined, the next steps involve model training, testing, and evaluation. During model training, the model learns to map input data to output predictions by adjusting its parameters through iterative optimization algorithms such as gradient descent. After training, the model is tested on a separate dataset to assess its performance on unseen data. Model evaluation metrics such as accuracy, precision, recall, and F1-score are calculated to quantify the model's performance. Finally, hyperparameter tuning is conducted to optimize the model's performance further by fine-tuning parameters such as learning rate, batch size, and dropout rate. This iterative process of training, testing, evaluation, and tuning continues until an optimized model with satisfactory performance is obtained. The model will be used to predict some knee joint images on the severity of OA from 0 to 4 based on the Kellgren-Lawrence grading system.

3.2.3 Model Architecture Involved

EfficientNet-B0

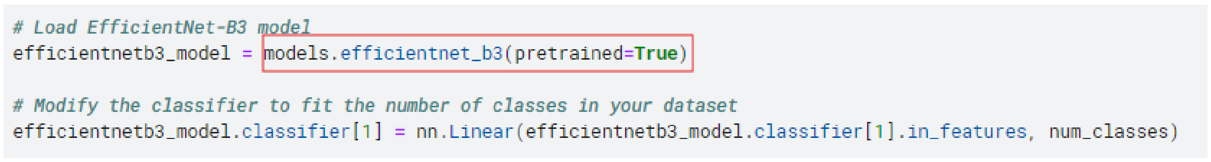


```
# Load EfficientNet-B0 model
efficientnet_model = models.efficientnet_b0(pretrained=True)

# Modify the classifier to fit the number of classes in your dataset
efficientnet_model.classifier[1] = nn.Linear(efficientnet_model.classifier[1].in_features, num_classes)
```

Figure 3.2.3 Code for EfficientNet-B0 adaptation

EfficientNet-B0 is a highly efficient baseline model that balances accuracy and computational cost, making it a strong candidate for your knee osteoarthritis grading task. Its compound scaling method ensures that the network scales depth, width, and resolution uniformly, optimizing performance without excessive computational demands. This can be particularly useful for working with a moderate-sized dataset of 2500 knee X-ray images, as it maintains high accuracy while being relatively resource-efficient.

EfficientNet-B3

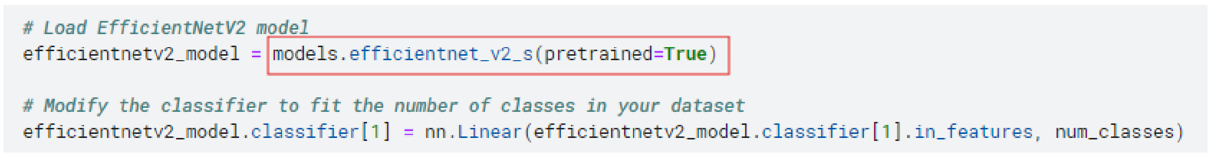


```
# Load EfficientNet-B3 model
efficientnetb3_model = models.efficientnet_b3(pretrained=True)

# Modify the classifier to fit the number of classes in your dataset
efficientnetb3_model.classifier[1] = nn.Linear(efficientnetb3_model.classifier[1].in_features, num_classes)
```

Figure 3.2.4 Code for EfficientNet-B3 adaptation

Building on the principles of EfficientNet-B0, EfficientNet-B3 offers increased capacity through additional depth, width, and resolution. This enhanced model can capture more complex patterns and features in your knee X-ray images, potentially leading to better performance in grading osteoarthritis. Its advanced scaling method ensures that improvements in accuracy come with manageable computational overhead, making it suitable for your dataset.

EfficientNetV2

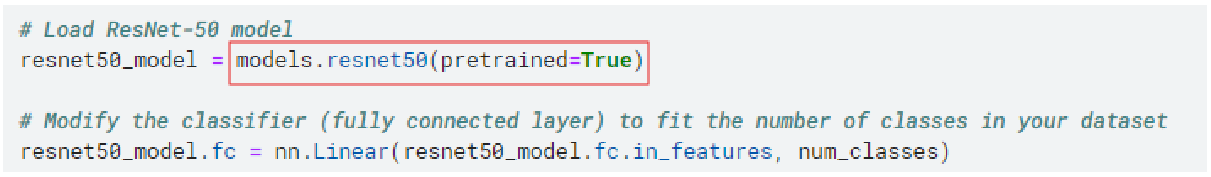


```
# Load EfficientNetV2 model
efficientnetv2_model = models.efficientnet_v2_s(pretrained=True)

# Modify the classifier to fit the number of classes in your dataset
efficientnetv2_model.classifier[1] = nn.Linear(efficientnetv2_model.classifier[1].in_features, num_classes)
```

Figure 3.2.5 Code for EfficientNetV2 adaptation

EfficientNetV2 introduces refined architecture and training methods that enhance both performance and efficiency over its predecessors. With its improved depthwise separable convolutions and compound scaling approach, EfficientNetV2 can handle the complexities of grading osteoarthritis from knee X-rays more effectively. Its efficiency makes it a good choice for managing your dataset size while achieving high accuracy.

ResNet50

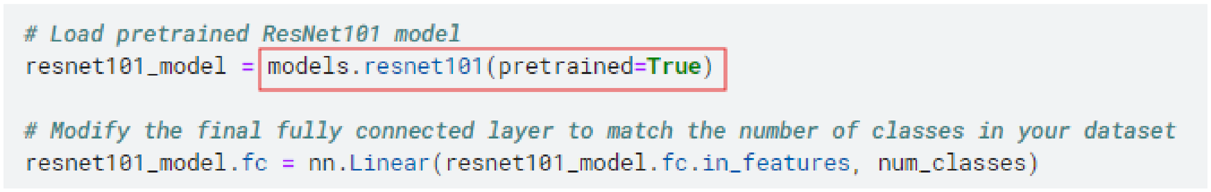


```
# Load ResNet-50 model
resnet50_model = models.resnet50(pretrained=True)

# Modify the classifier (fully connected layer) to fit the number of classes in your dataset
resnet50_model.fc = nn.Linear(resnet50_model.fc.in_features, num_classes)
```

Figure 3.2.6 Code for ResNet50 adaptation

ResNet50's residual connections help mitigate the vanishing gradient problem, allowing it to learn more complex features from knee X-ray images. Its 50 layers make it deep enough to capture intricate details in the images, which is crucial for accurate osteoarthritis grading. The residual learning framework supports effective training and convergence, making ResNet50 a reliable choice for your task.

ResNet101

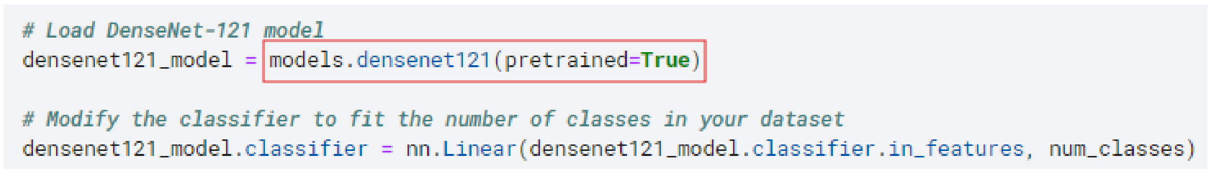


```
# Load pretrained ResNet101 model
resnet101_model = models.resnet101(pretrained=True)

# Modify the final fully connected layer to match the number of classes in your dataset
resnet101_model.fc = nn.Linear(resnet101_model.fc.in_features, num_classes)
```

Figure 3.2.7 Code for ResNet101 adaptation

Extending ResNet50, ResNet101 offers even greater depth, enabling it to learn more nuanced features from knee X-ray images. This deeper architecture improves the model's ability to detect subtle changes in bone structures indicative of osteoarthritis. The residual connections continue to facilitate effective training, making ResNet101 a powerful option for detailed grading tasks.

DenseNet121

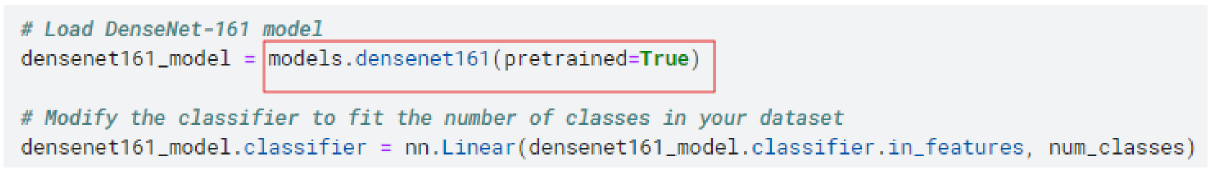


```python
# Load DenseNet-121 model
densenet121_model = models.densenet121(pretrained=True)

# Modify the classifier to fit the number of classes in your dataset
densenet121_model.classifier = nn.Linear(densenet121_model.classifier.in_features, num_classes)
```

Figure 3.2.8 Code for DenseNet121 adaptation

DenseNet121's dense connectivity pattern allows each layer to access features from all previous layers, promoting feature reuse and enhancing gradient flow. This architecture is well-suited for capturing detailed and complex patterns in knee X-ray images, which can improve the accuracy of osteoarthritis grading while reducing the number of parameters needed.

DenseNet161

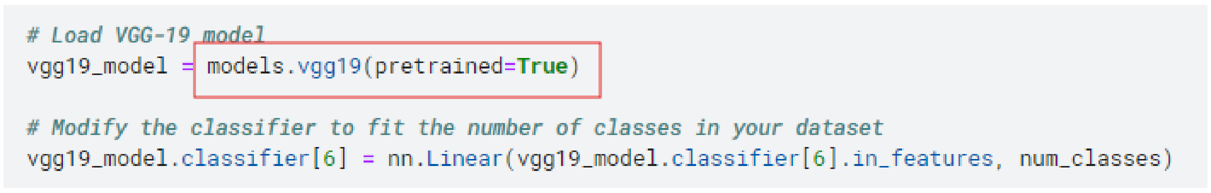


```python
# Load DenseNet-161 model
densenet161_model = models.densenet161(pretrained=True)

# Modify the classifier to fit the number of classes in your dataset
densenet161_model.classifier = nn.Linear(densenet161_model.classifier.in_features, num_classes)
```

Figure 3.2.9 Code for DenseNet161 adaptation

With an increased number of layers, DenseNet161 provides enhanced feature extraction capabilities, which can be beneficial for grading osteoarthritis from knee X-rays. Its dense connectivity further supports effective gradient flow and feature reuse, helping the model to identify subtle changes in bone structures more accurately.

VGG-19

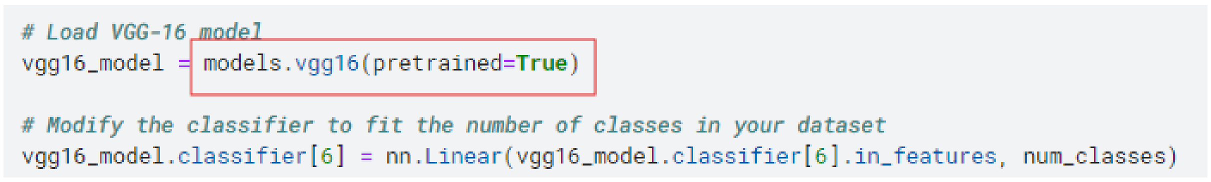


```
# Load VGG-19 model
vgg19_model = models.vgg19(pretrained=True)

# Modify the classifier to fit the number of classes in your dataset
vgg19_model.classifier[6] = nn.Linear(vgg19_model.classifier[6].in_features, num_classes)
```

Figure 3.2.10 Code for VGG-19 adaptation

VGG-19's deep architecture and use of small convolutional filters make it effective at capturing detailed features in knee X-ray images. Its straightforward design and effectiveness in feature extraction can help in grading osteoarthritis by identifying key visual indicators in the images. Despite its depth, VGG-19 remains manageable for your dataset size.

VGG-16

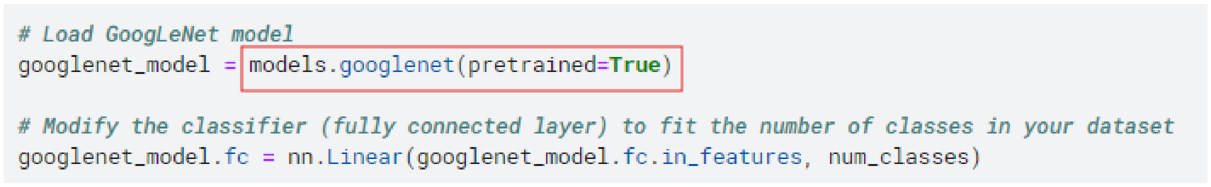


```
# Load VGG-16 model
vgg16_model = models.vgg16(pretrained=True)

# Modify the classifier to fit the number of classes in your dataset
vgg16_model.classifier[6] = nn.Linear(vgg16_model.classifier[6].in_features, num_classes)
```

Figure 3.2.11 Code for VGG-16 adaptation

VGG-16, with 16 layers, shares similar architectural principles with VGG-19 but is slightly less deep. It is still effective for feature extraction in knee X-ray images and can provide reliable grading results for osteoarthritis. Its simpler architecture may also result in faster training times compared to deeper models.

GoogLeNet



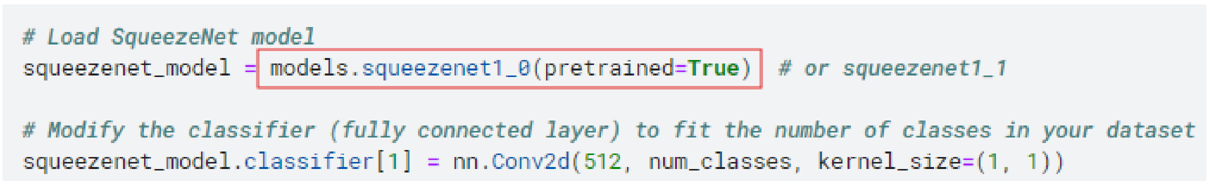
```
# Load GoogLeNet model
googlenet_model = models.googlenet(pretrained=True)

# Modify the classifier (fully connected layer) to fit the number of classes in your dataset
googlenet_model.fc = nn.Linear(googlenet_model.fc.in_features, num_classes)
```

Figure 3.2.12 Code for GoogLeNet adaptation

GoogLeNet's Inception modules allow it to capture features at multiple scales within a single layer, which can be beneficial for analyzing complex patterns in knee X-ray images.

This multi-scale approach helps in extracting a rich set of features for accurate osteoarthritis grading, while its efficient design reduces computational costs.

SqueezeNet

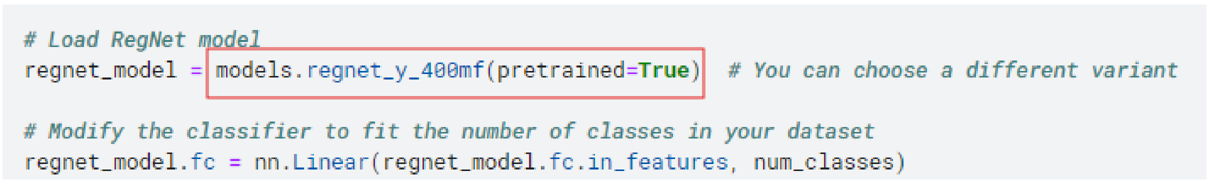


```
# Load SqueezeNet model
squeezenet_model = models.squeezenet1_0(pretrained=True)  # or squeezenet1_1

# Modify the classifier (fully connected layer) to fit the number of classes in your dataset
squeezenet_model.classifier[1] = nn.Conv2d(512, num_classes, kernel_size=(1, 1))
```

Figure 3.2.13 Code for SqueezeNet adaptation

SqueezeNet is designed for efficiency with a smaller number of parameters, making it suitable for resource-constrained environments. Its fire modules balance performance with efficiency, allowing it to work well with a dataset of 2500 knee X-ray images while still providing good accuracy for osteoarthritis grading.

RegNet

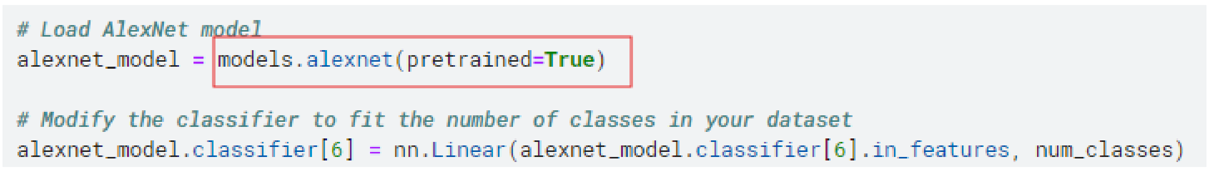


```
# Load RegNet model
regnet_model = models.regnet_y_400mf(pretrained=True)  # You can choose a different variant

# Modify the classifier to fit the number of classes in your dataset
regnet_model.fc = nn.Linear(regnet_model.fc.in_features, num_classes)
```

Figure 3.2.14 Code for RegNet adaptation

RegNet's scalable and flexible architecture allows it to be easily adjusted for different computational resources. Its regularized design can efficiently handle the feature extraction needs for knee X-ray images, making it a good choice for your osteoarthritis grading task due to its adaptability and balanced performance.

AlexNet



```
# Load AlexNet model
alexnet_model = models.alexnet(pretrained=True)

# Modify the classifier to fit the number of classes in your dataset
alexnet_model.classifier[6] = nn.Linear(alexnet_model.classifier[6].in_features, num_classes)
```

Figure 3.2.15 Code for AlexNet adaptation

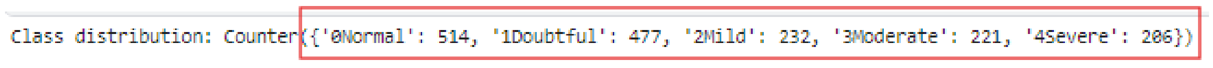
As a pioneering deep learning model, AlexNet's deep convolutional architecture and techniques like ReLU activations and dropout regularization laid the groundwork for modern image classification. Although it is not as advanced as newer models, its simplicity and effectiveness in extracting features from knee X-ray images can still be valuable for grading osteoarthritis, especially if computational resources are limited.

Chapter 4 System Implementation

4.1 Data Preprocessing

4.1.1 Original Dataset

Upon inspection, the dataset imposes noticeable imbalance in class distribution. Class 0 (Normal) contained 514 images, while class 4 (Severe) had only 206 images. This imbalance posed significant difficulties during model training, as the model tended to favour the majority classes, leading to biased predictions.



```
Class distribution: Counter({'0Normal': 514, '1Doubtful': 477, '2Mild': 232, '3Moderate': 221, '4Severe': 206})
```

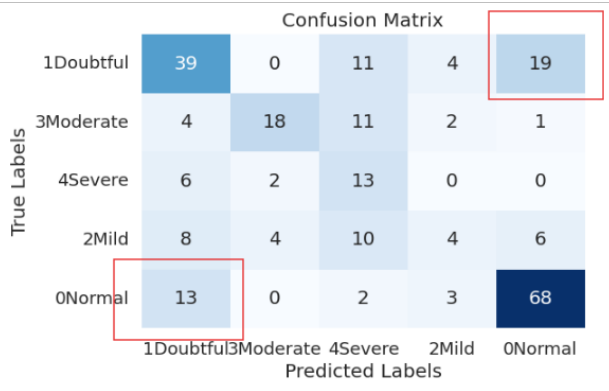
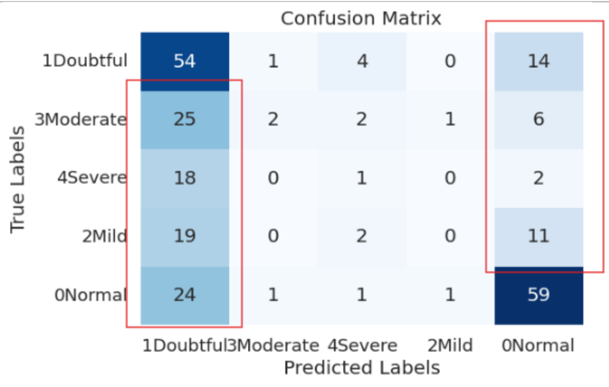
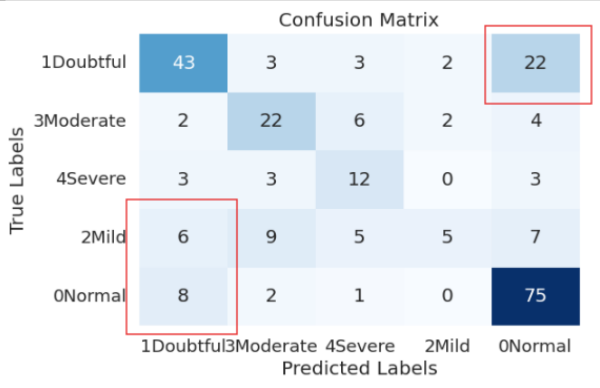
Figure 4.1.1 Inspected imbalance class size in dataset

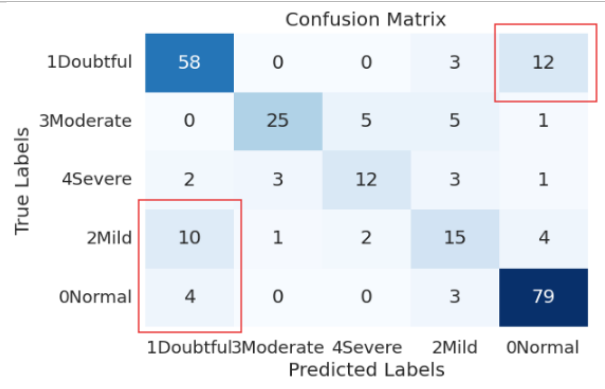
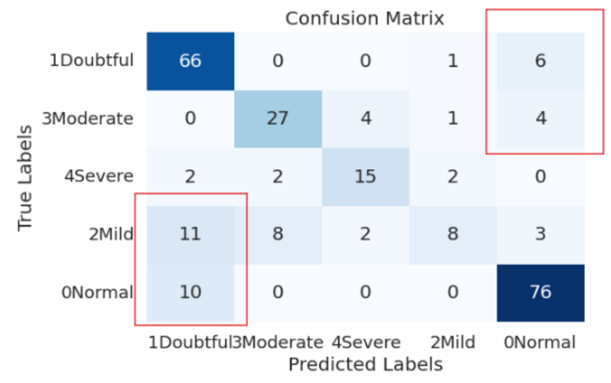
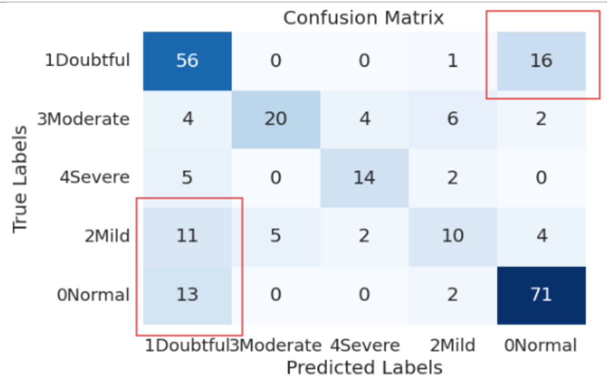
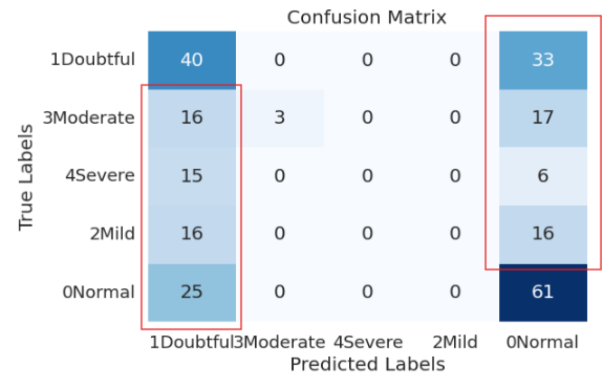
Table 4.1 Testing accuracy and F1 score for model trained with imbalance dataset

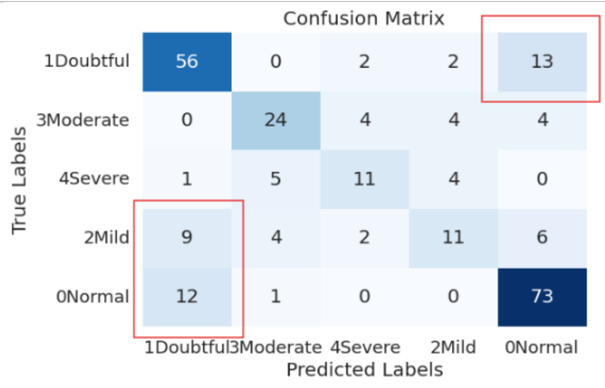
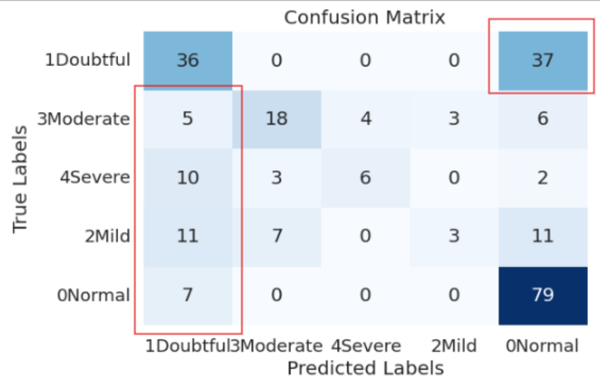
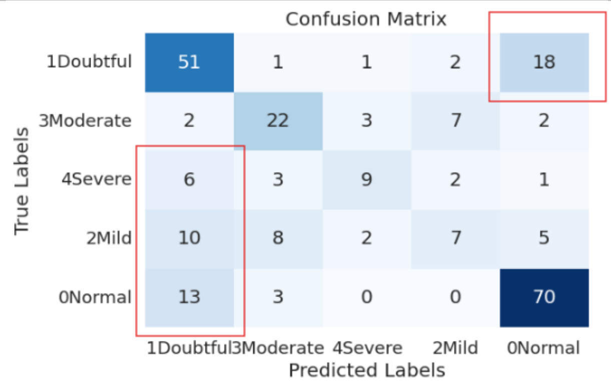
Architecture	*Accuracy*	*F1-Score*
EfficientNet-B0	57.26	49.22
EfficientNet-B3	46.77	26.69
EfficientNetV2	63.31	54.58
ResNet50	**76.21**	70.27
ResNet101	**77.42**	70.02
DenseNet121	68.95	64.01
DenseNet161	41.94	22.87
VGG-19	**70.56**	63.89
GoogleNet	57.26	46.59
SqueezeNet	64.11	56.20

When training the model on the original, unbalanced dataset, it became clear that achieving high accuracy was challenging. Despite multiple attempts at optimizing the model and try to train more variety of model, the accuracy never exceeded 80%. Furthermore, the class imbalance resulted in an inflated false positive rate for classes with a larger number of samples, particularly for the 0 (Normal) and 1 (Doubtful) categories. The model frequently misclassified uncertain images into these larger classes, as the higher frequency of these labels increased the probability of correct predictions. While this increased the overall accuracy, it came at the cost of AUC (Area Under the Curve), which remained poor. AUC, which measures the ability of the model to distinguish between classes, reflected the model's failure to correctly identify minority class instances such as those in the 2 (Mild), 3 (Moderate) and 4 (Severe) categories. As a result, although accuracy seemed promising, it was not a reliable indicator of model performance in this case.

Table 4.2 Confusion matrix for each model trained with imbalance dataset

Architecture	**Confusion Matrix**
EfficientNet-B0	
EfficientNet-B3	
EfficientNetV2	

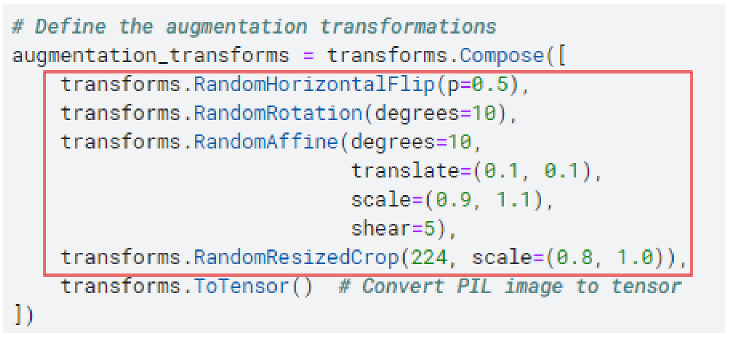
ResNet50	
ResNet101	
DenseNet121	
DenseNet161	

VGG-19	
GoogLeNet	
SqueezeNet	

Across all models tested, including ResNet50, ResNet101, and VGG-19, which achieved over 70% accuracy, it was observed that classes 0 (Normal) and 1 (Doubtful) consistently exhibited the highest false positive rates. This trend was particularly pronounced due to the fact that these classes had the largest number of data points, leading the models to disproportionately favour them when making predictions. Even for well-performing models, the imbalance in data distribution made it difficult to accurately classify the minority classes, causing the models to misclassify many instances from the less frequent classes as belonging to class 0 or 1. This issue was further exacerbated in models like EfficientNetB0, EfficientNetB3, DenseNet161, and GoogLeNet, where a significant portion of the data was almost automatically classified into class 0 or 1, severely limiting the models' ability to generalize beyond these two categories. This behaviour can be attributed to the models optimizing for overall accuracy, which in an imbalanced dataset, results in the model focusing on getting the most frequent classes correct at the expense of minority classes. As a result, although accuracy appeared high or moderate, the models' actual performance on distinguishing between classes was poor, particularly when evaluated using metrics like the F1 score and AUC. These findings emphasize the importance of addressing class imbalance when training deep learning models, as high accuracy alone is not a reliable indicator of a model's true predictive power.

4.1.2 Balanced Dataset

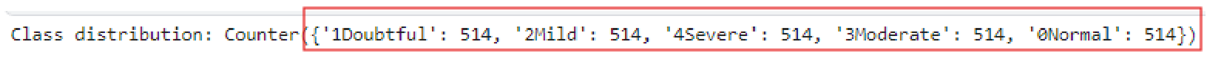
To address the imbalanced dataset problem faced, several data augmentation techniques were implemented to balance the dataset and improve model generalization ability. The augmentation technique used to balance the dataset applies several random transformations to increase the diversity of training images, particularly in the minority classes. These transformations include random horizontal flipping, which provides different orientations, and random rotation, adding variety in angles to make the model more robust to slight misalignments. Affine transformations are also applied, including translation (shifting images), scaling (adjusting image size), and shearing (skewing), which simulate various real-world conditions such as different object placements and sizes in X-rays. Additionally, random resized cropping alters the image composition by varying the area cropped and resizing it to a consistent 224x224 pixels. These techniques not only create a more balanced dataset but also enhance the model's ability to generalize by exposing it to a wider range of image variations.



```
# Define the augmentation transformations
augmentation_transforms = transforms.Compose([
    transforms.RandomHorizontalFlip(p=0.5),
    transforms.RandomRotation(degrees=10),
    transforms.RandomAffine(degrees=10,
                            translate=(0.1, 0.1),
                            scale=(0.9, 1.1),
                            shear=5),
    transforms.RandomResizedCrop(224, scale=(0.8, 1.0)),
    transforms.ToTensor()  # Convert PIL image to tensor
])
```

Figure 4.1.2 Code for data augmentation to balance dataset

After performing the data augmentation techniques, amount of data in all 5 classes was even out with each containing 514 images.



```
Class distribution: Counter({'1Doubtful': 514, '2Mild': 514, '4Severe': 514, '3Moderate': 514, '0Normal': 514})
```

Figure 4.1.3 Inspected balanced dataset after data augmentation

Table 4.3 Comparison of testing accuracy for model before and after balancing

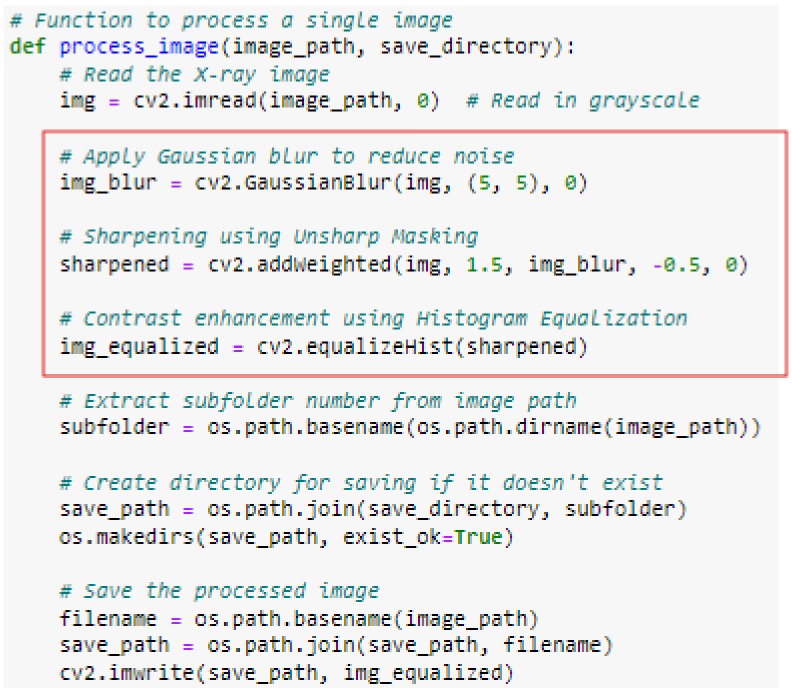
Architecture	*Accuracy*		*F1 Score*
	Previous	Current	
EfficientNet-B0	57.26	68.91	67.33
EfficientNet-B3	46.77	58.55	54.03
EfficientNetV2	63.31	75.91	74.36
ResNet50	76.21	**80.57**	80.07
ResNet101	77.42	**87.31**	86.82
DenseNet121	68.95	79.27	78.78
DenseNet161	41.94	58.29	55.00
VGG-19	70.56	76.42	75.55
VGG-16	-	76.94	76.15
GoogleNet	57.26	71.50	70.27
SqueezeNet	64.11	68.65	67.42
RegNet	-	76.94	76.75
AlexNet	-	67.36	65.40

After applying the augmentation techniques, all of the models showed significant improvement in both accuracy and F1 score, addressing the earlier issue of misclassifying most images into classes 0 and 1. With the augmented dataset, the models became more adept at distinguishing between the different severity levels of osteoarthritis. Notably, ResNet50 and ResNet101 achieved accuracies of 80.57% and 87.31%, respectively, showcasing their effectiveness in handling the newly balanced dataset.

In addition to these, three more models—VGG-19, RegNet, and AlexNet, were trained. Among all models trained, nearly all of them achieved over 70% accuracy and F1 score. The improvements in F1 score indicate that not only did the models become more accurate, but they also improved their ability to handle both false positives and false negatives across all classes. This demonstrates that the augmentation technique successfully balanced the dataset, allowing the models to learn more effectively from both majority and minority classes. The reduction in bias towards classes 0 and 1, combined with the enhanced generalization provided by the data augmentation, resulted in more reliable and consistent performance across all the models tested.

4.1.3 Sharpened Dataset

After addressing the dataset balancing issue, an additional improvement was made by applying an image sharpening technique using OpenCV to enhance the dataset further. This technique involves several key steps: First, Gaussian blur is applied to the X-ray images to reduce noise and smooth out minor variations. Next, sharpening is performed using unsharp masking, which enhances the image details by subtracting a blurred version of the image from the original, effectively making edges more distinct. Finally, contrast enhancement is achieved through histogram equalization, which improves the visibility of features by adjusting the image's intensity distribution. The processed images are then saved into a structured directory for consistency. This sharpening technique aims to enhance the clarity of subtle details in the X-ray images, making it easier for the model to detect and differentiate between various levels of osteoarthritis severity. By improving the image quality, the model's ability to accurately classify the severity of the condition is further refined.



```python
# Function to process a single image
def process_image(image_path, save_directory):
    # Read the X-ray image
    img = cv2.imread(image_path, 0)  # Read in grayscale

    # Apply Gaussian blur to reduce noise
    img_blur = cv2.GaussianBlur(img, (5, 5), 0)

    # Sharpening using Unsharp Masking
    sharpened = cv2.addWeighted(img, 1.5, img_blur, -0.5, 0)

    # Contrast enhancement using Histogram Equalization
    img_equalized = cv2.equalizeHist(sharpened)

    # Extract subfolder number from image path
    subfolder = os.path.basename(os.path.dirname(image_path))

    # Create directory for saving if it doesn't exist
    save_path = os.path.join(save_directory, subfolder)
    os.makedirs(save_path, exist_ok=True)

    # Save the processed image
    filename = os.path.basename(image_path)
    save_path = os.path.join(save_path, filename)
    cv2.imwrite(save_path, img_equalized)
```

Figure 4.1.4 Code for image sharpening

Table 4.4 Comparison for testing accuracy of model before and after image sharpening

Architecture	*Accuracy*	
	Previous	Current
EfficientNet-B0	68.91	68.65
EfficientNet-B3	58.55	58.03
EfficientNetV2	75.91	76.68
ResNet50	**80.57**	**82.90**
ResNet101	**87.31**	**81.87**
DenseNet121	79.27	**80.83**
DenseNet161	58.29	61.66
VGG-19	76.42	75.65
VGG-16	76.94	77.72
GoogleNet	71.50	72.02
SqueezeNet	68.65	74.09
RegNet	76.94	76.17
AlexNet	67.36	73.58

After applying the sharpening technique, the model performance showed slight improvements, with ResNet50, ResNet101, and DenseNet121 achieving accuracies over 80%. However, the most notable improvements were observed with SqueezeNet and AlexNet, which saw nearly a 5% increase in accuracy. SqueezeNet and AlexNet experienced more significant gains because of their unique architectures and their sensitivity to image details. SqueezeNet, with its compact design, relies heavily on efficient feature extraction, making it particularly responsive to enhancements in image sharpness, which improves its ability to discern fine details. AlexNet, on the other hand, is known for its deeper layers and larger receptive fields, which can benefit substantially from improved image clarity as it allows the network to better capture and utilize high-frequency features. Both models are more affected by changes in image quality compared to deeper, more complex networks like ResNet and DenseNet, which have more advanced mechanisms for feature extraction and noise reduction built into their architectures. This heightened sensitivity to image quality improvements explains why SqueezeNet and AlexNet showed more pronounced performance enhancements.

4.2 Model Training

Learning Curve & Testing Accuracy

Training Settings:

Learning Rate = 0.001

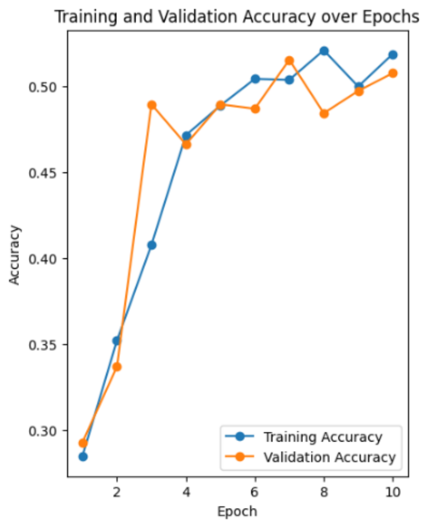
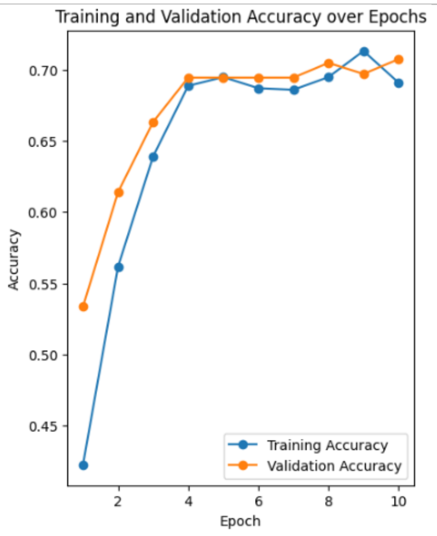
Batch Size = 16

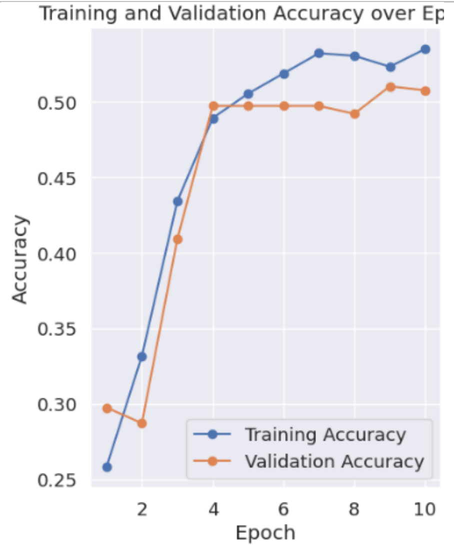
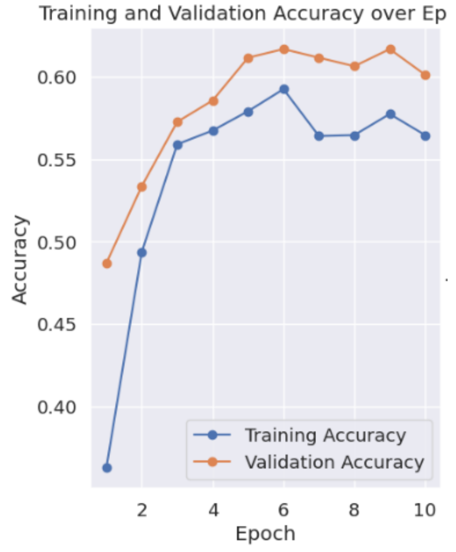
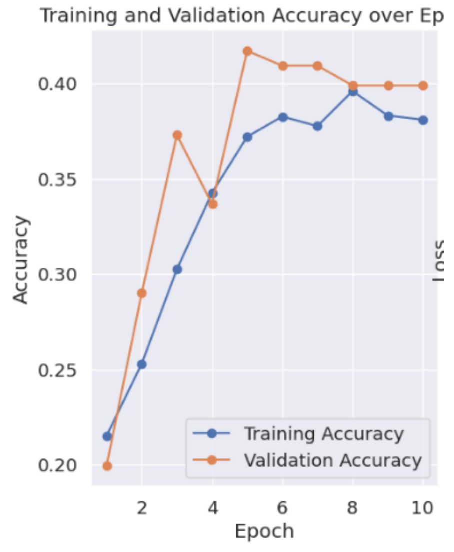
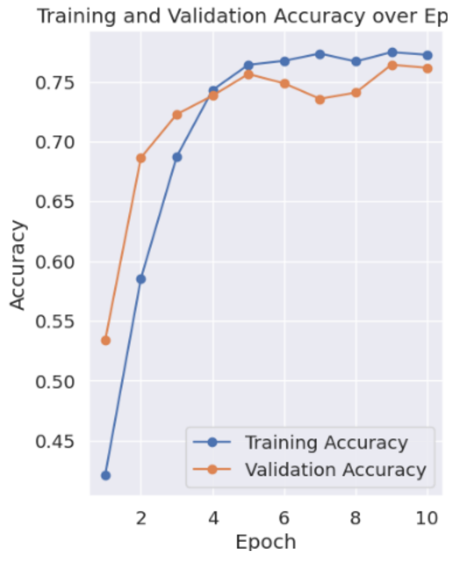
Optimizer = Stochastic Gradient Descent (SGD)

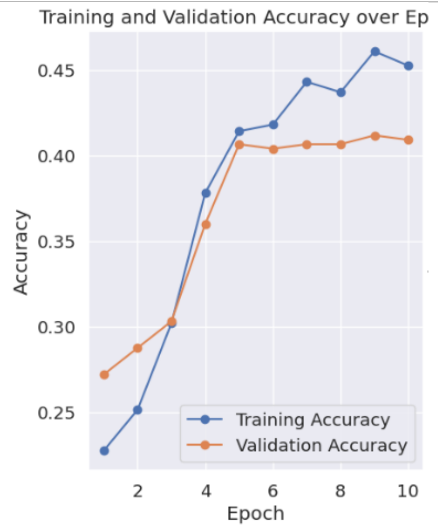
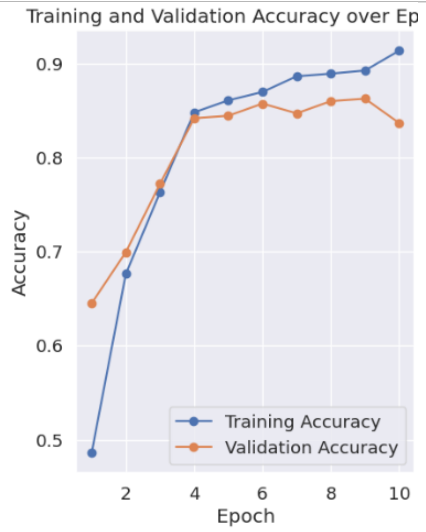
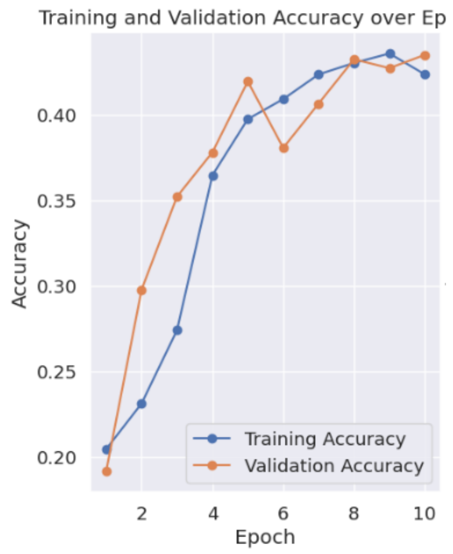
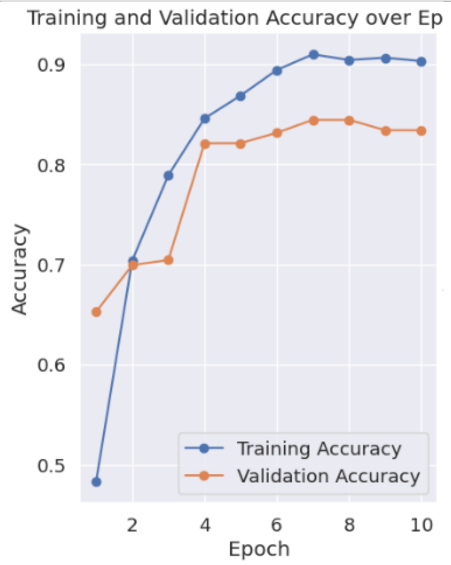
Epoch = 10

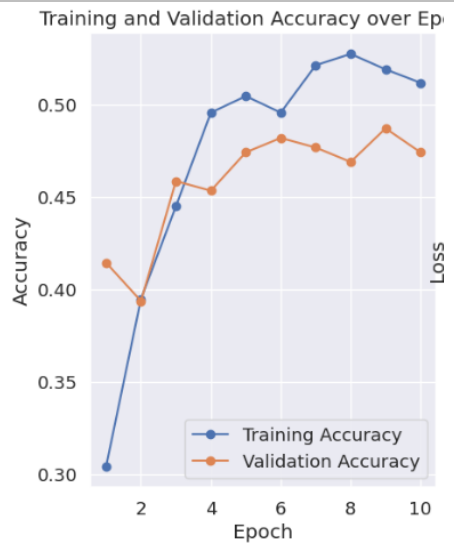
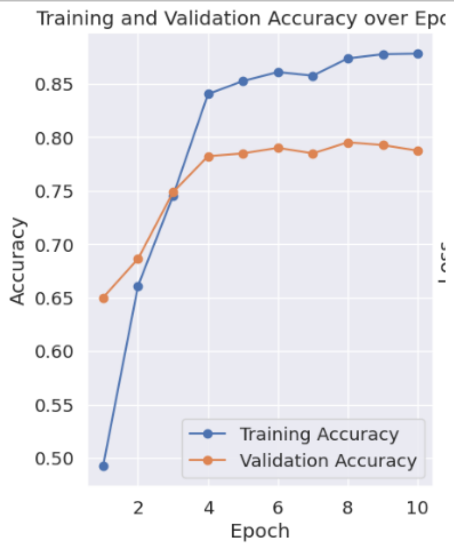
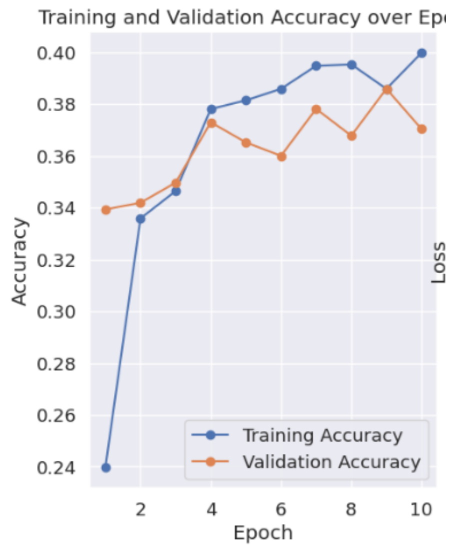
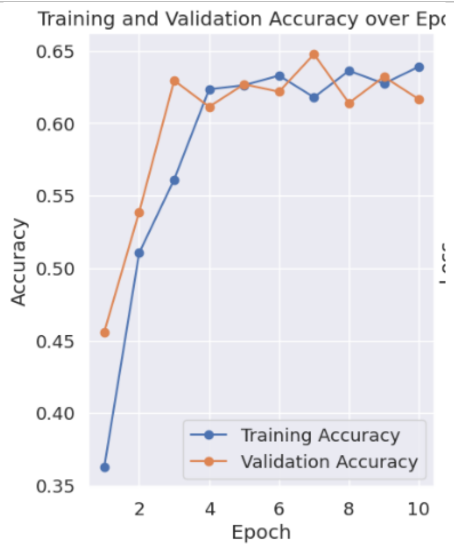
Dataset = Balanced & Image Sharpened Dataset

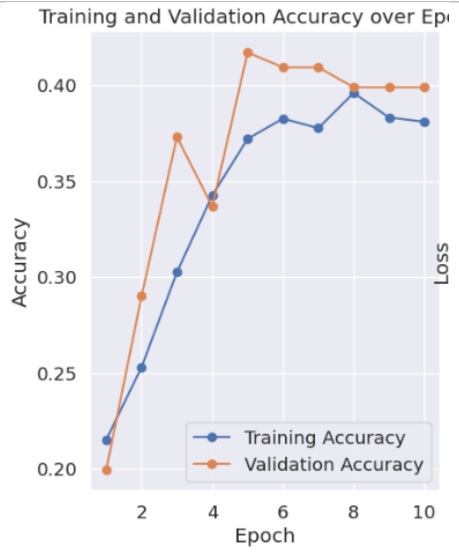
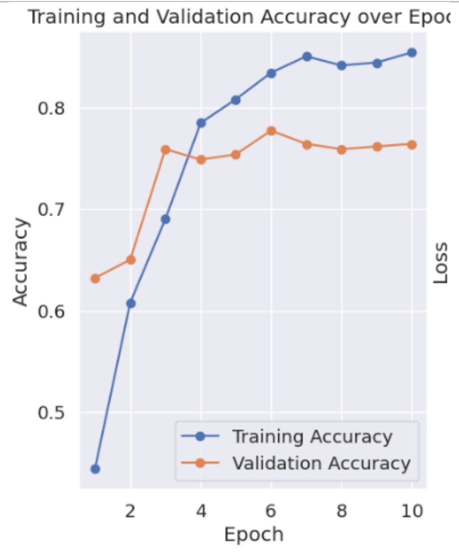
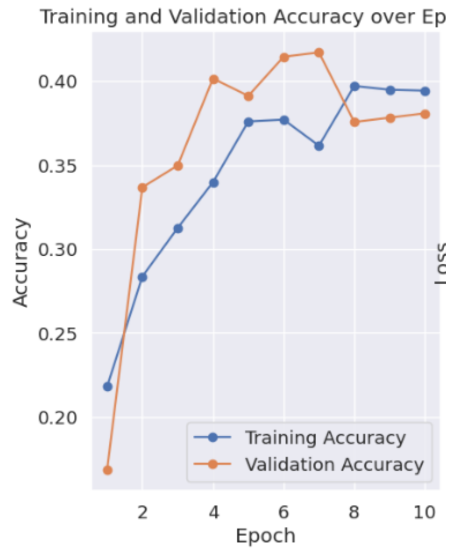
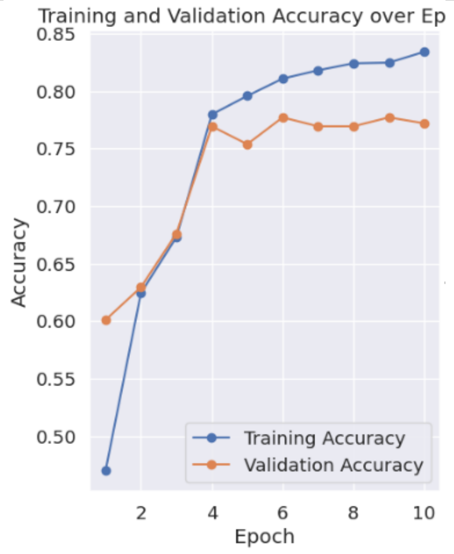
Table 4.5 Comparison for original and pretrained model

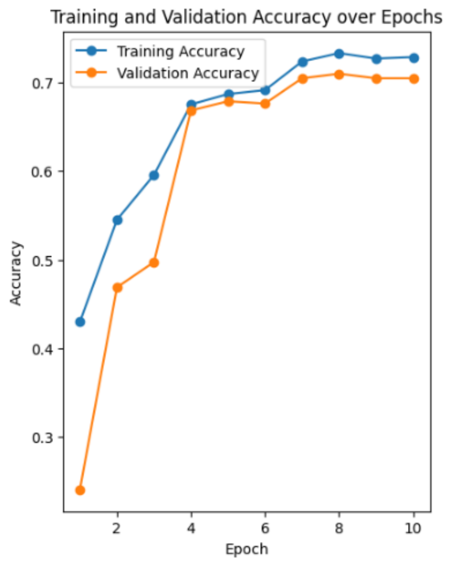
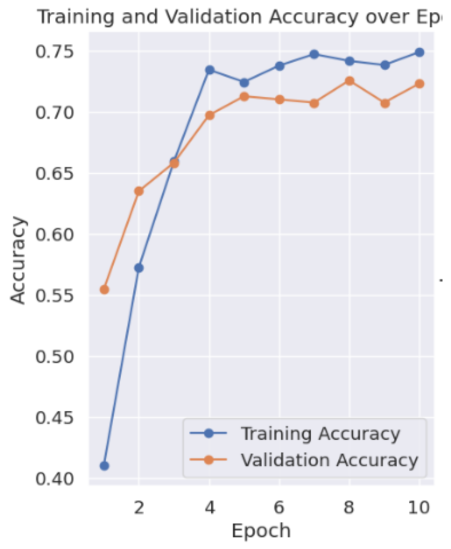
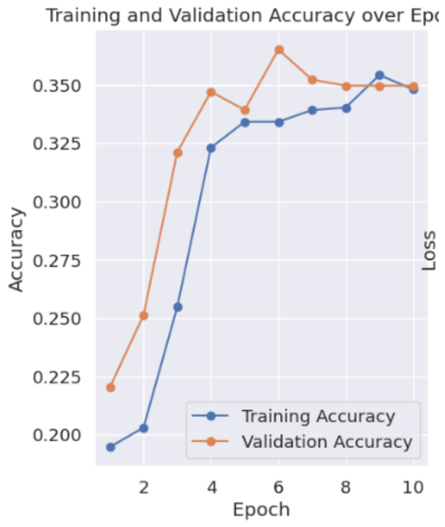
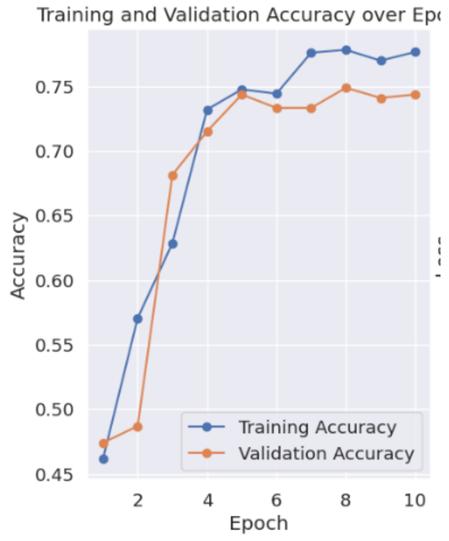
Architecture	Original Model		Pretrained Model	
	Training History	Testing Accuracy (%)	Training History	Testing Accuracy (%)
EfficientNet-B0		52.85		68.65

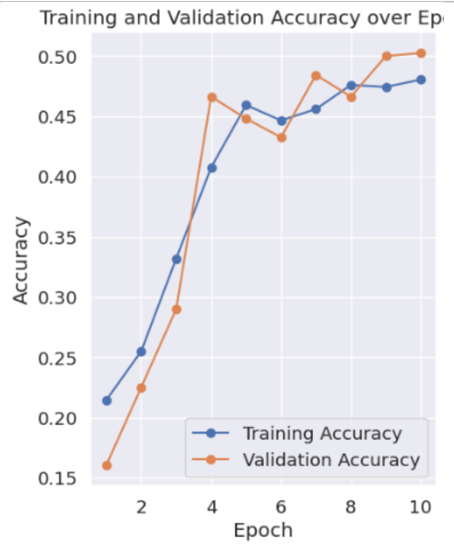
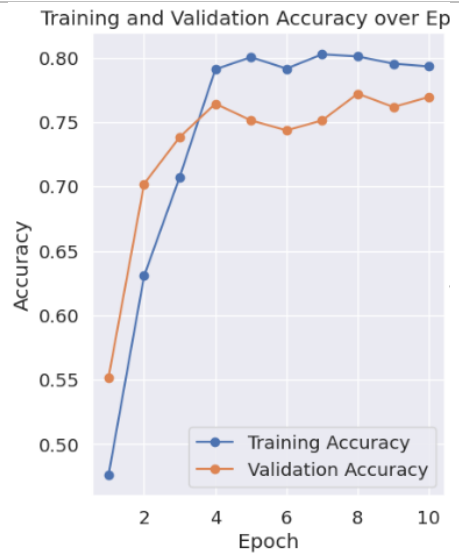
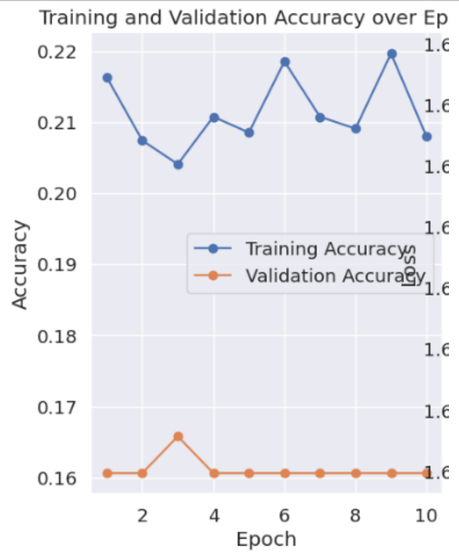
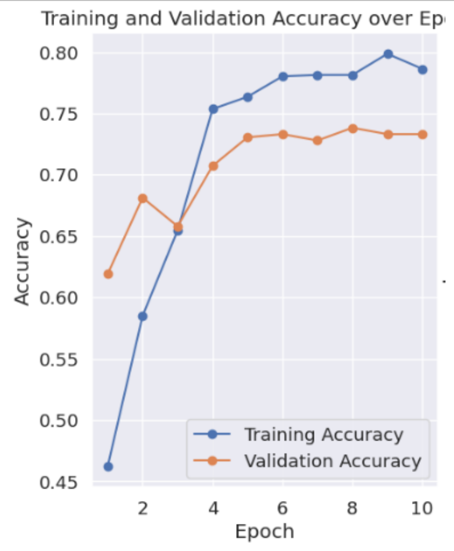
EfficientNet-B3		55.70		58.03
EfficientNetV2		38.86		76.68

ResNet50		41.97		**82.90**
ResNet101		42.49		**81.87**

DenseNet121		47.67		**80.83**
DenseNet161		39.90		61.66

VGG-19		43.26		75.65
VGG-16		44.04		77.72

GoogleNet		68.91		72.02
SqueezeNet		35.75		74.09

RegNet		44.56		76.17
AlexNet		20.98		73.58

The observed performance differences between pretrained and original models in the knee osteoarthritis grading research can be largely attributed to the advantages of leveraging previously learned features. Pretrained models benefit from initial weights acquired from large-scale datasets such as ImageNet, which provide them with a robust foundation of visual features. This prior knowledge enables pretrained models to converge faster and achieve better generalization on new tasks, in this case, the knee osteoarthritis grading. In contrast, untrained models must learn all features from scratch, which significantly slows their performance improvement and results in higher validation loss and lower accuracy.

Models like ResNet50, ResNet101, and DenseNet121, which achieved over 80% accuracy when pretrained, experienced substantial improvements compared to their untrained versions, which remained below 50%. EfficientNet-B0 saw an increase from 52.85% to 68.65%, and EfficientNetV2 demonstrated an even larger boost from 38.86% to 76.68%. These improvements highlight how pretrained models, starting with a more comprehensive understanding of visual features, adapt more quickly and effectively to specific tasks.

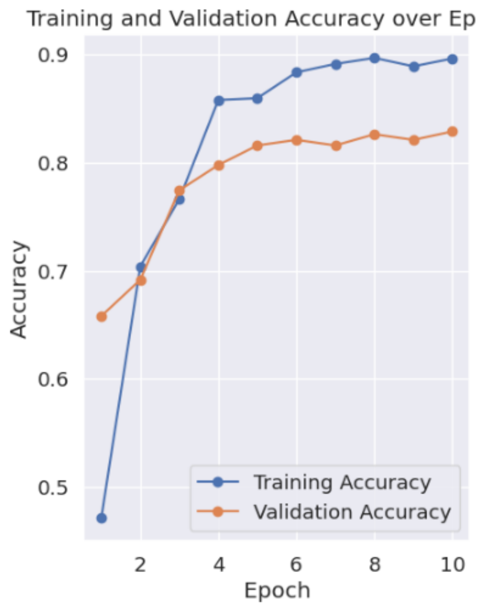
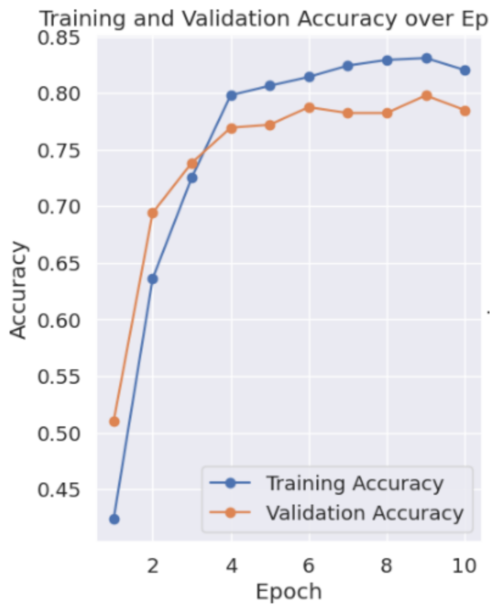
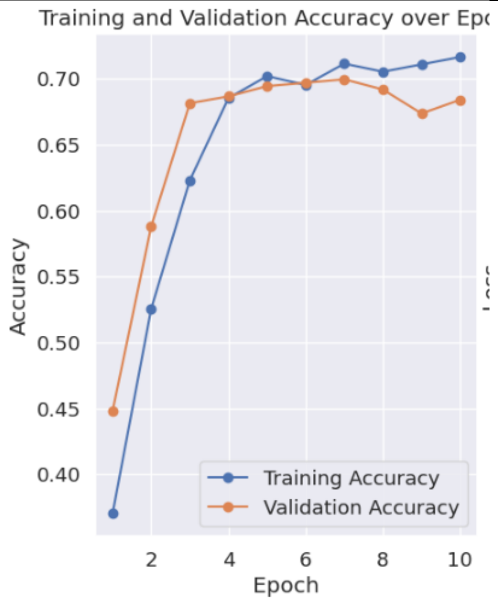
The architecture of the models also plays a significant role in performance. Deeper architectures, such as ResNet and DenseNet, which incorporate advanced feature extraction mechanisms like residual and dense connections, benefit significantly from pretraining. These models are better equipped to capture complex and hierarchical features, crucial for accurately grading knee osteoarthritis. Conversely, simpler models like AlexNet and SqueezeNet, while showing improvements, lag behind due to their less sophisticated feature extraction capabilities. This architectural variance explains why deeper models generally achieve higher performance, as they are more adept at learning from intricate patterns in the data.

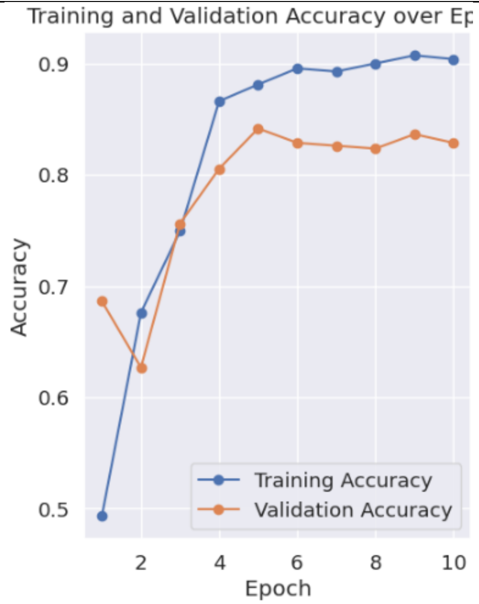
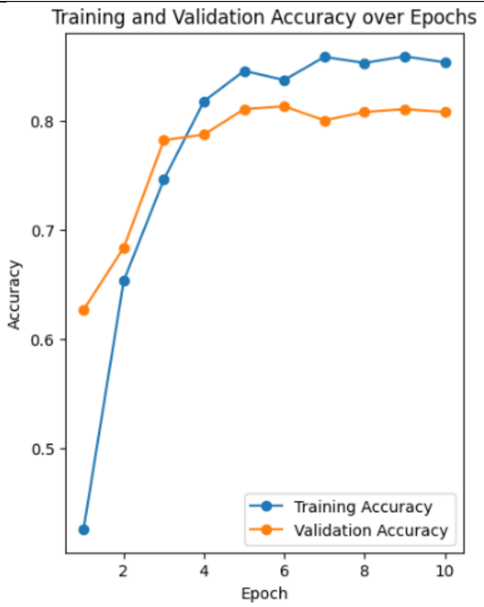
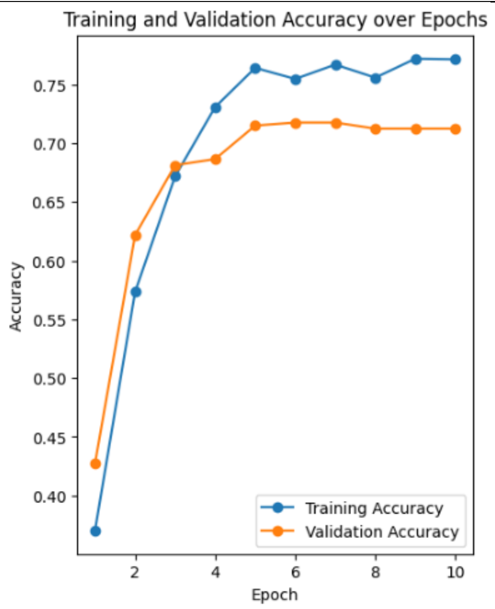
Among all of the models trained, ResNet50 achieved highest accuracy of 82.90% test accuracy closely followed by ResNet101 with test accuracy of 81.87%. The third is place is the DenseNet121 with 80.83% test accuracy. After inspecting the training history, the training accuracy and validation accuracy are near in the range as the testing accuracy indicating that possibility for model overfitting issue is small. These 3 models will proceed to the next part for hyperparameter fine-tuning to find more possibility in improving the performance.

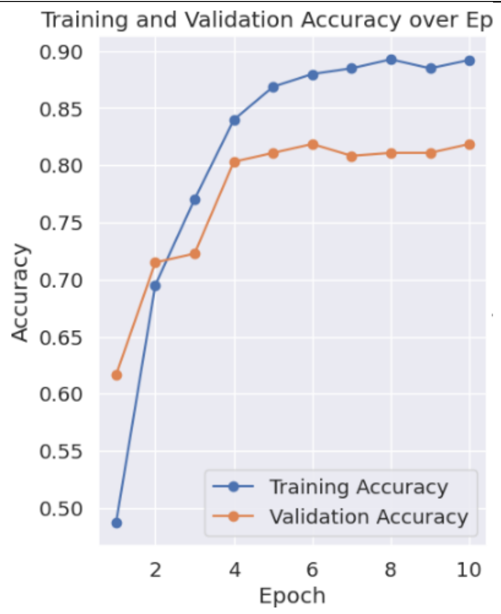
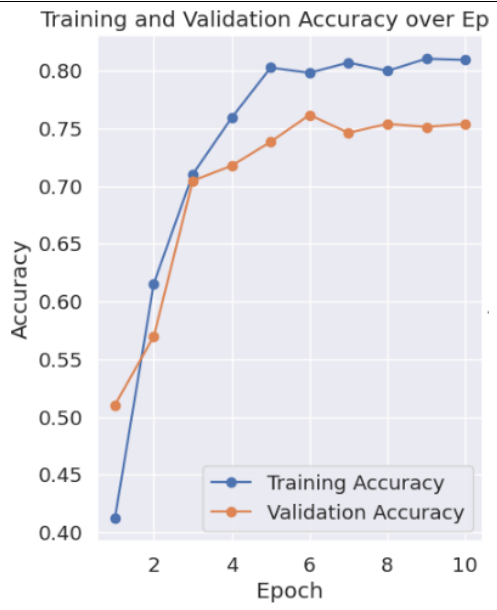
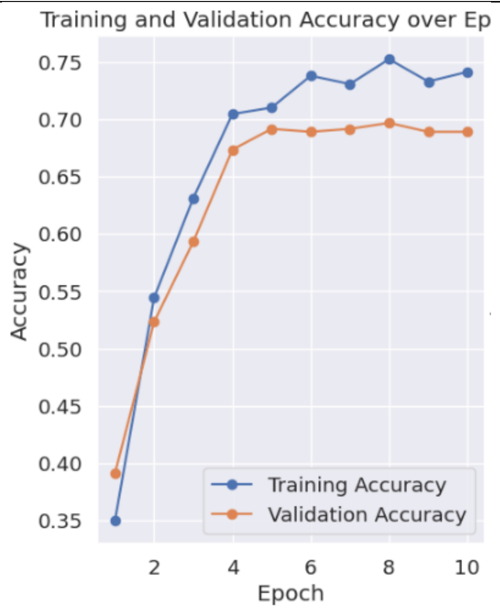
4.3 Hyperparameter Fine-Tuning

4.3.1 Batch Size

Table 4.6 Comparison for model performance with different batch size

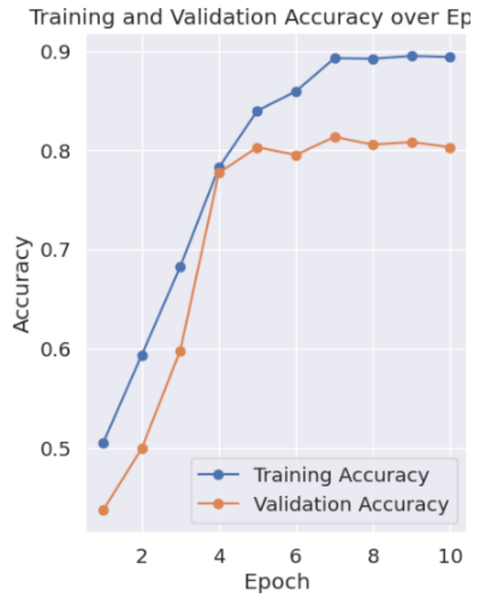
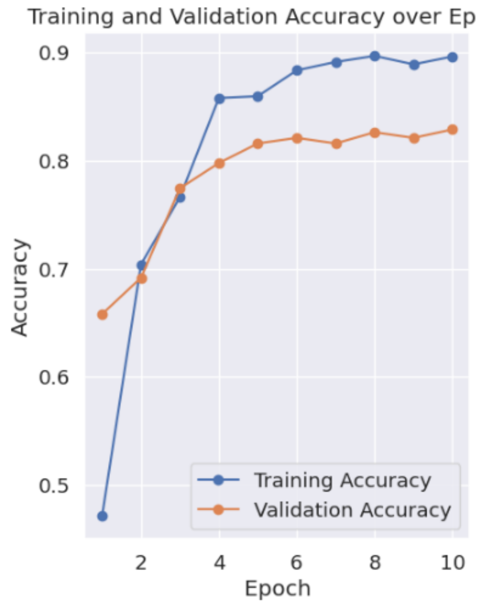
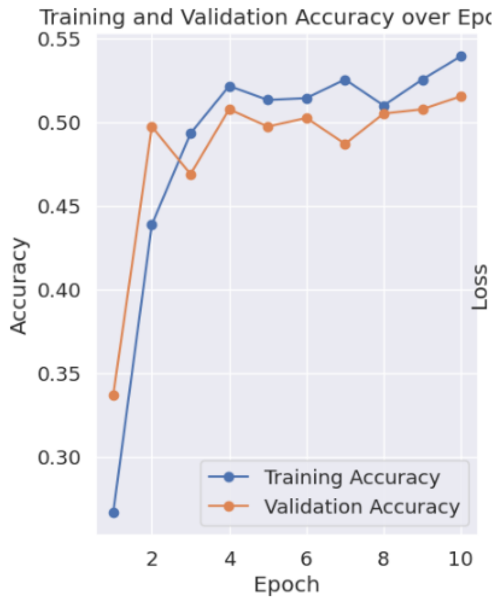
Architecture	Batch Size		
	16	32	64
ResNet50	 Training Accuracy: 89.66% Validation Accuracy: 82.90% Testing Accuracy: 83.94%	 Training Accuracy: 82.04% Validation Accuracy: 78.50% Testing Accuracy: 75.65%	 Training Accuracy: 71.64% Validation Accuracy: 68.39% Testing Accuracy: 65.03%

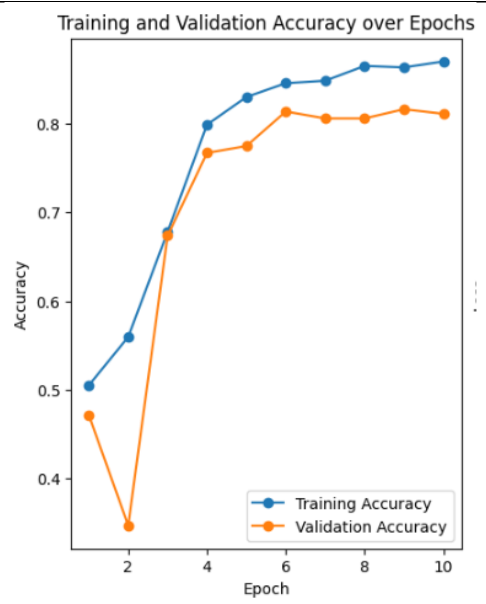
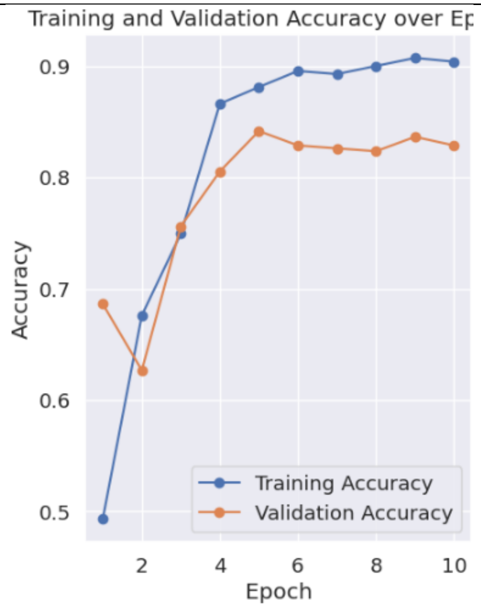
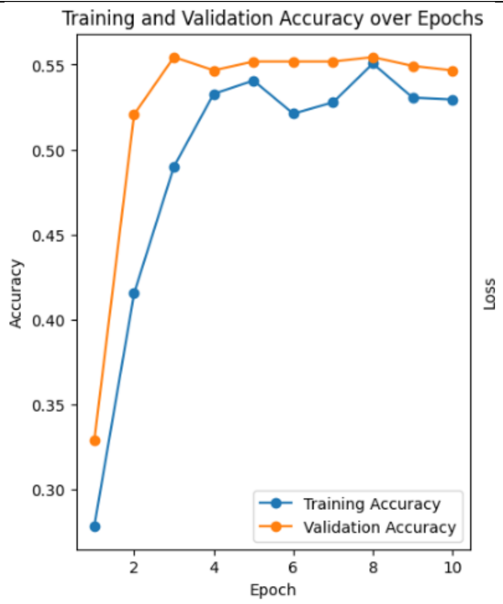
ResNet101	 Training Accuracy: 90.43% Validation Accuracy: 82.90% Testing Accuracy: 84.72%	 Training Accuracy: 85.37% Validation Accuracy: 80.83% Testing Accuracy: 75.39%	 Training Accuracy: 77.14% Validation Accuracy: 71.24% Testing Accuracy: 73.58%

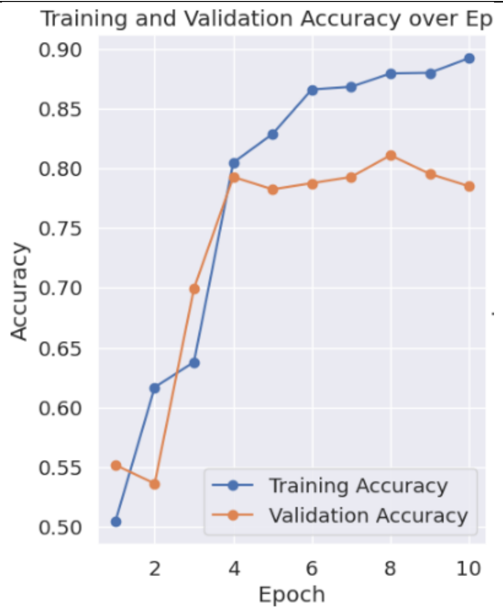
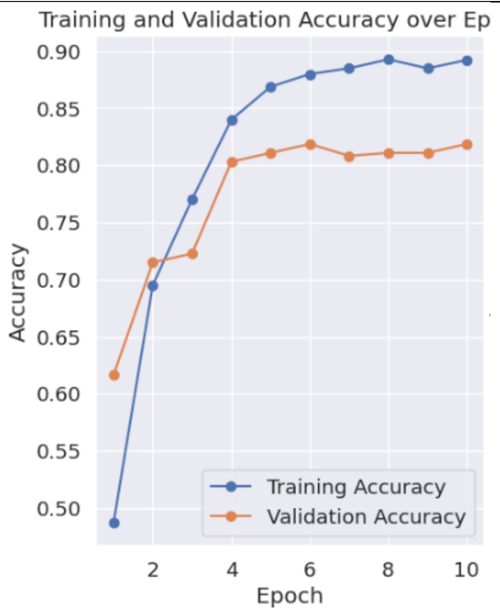
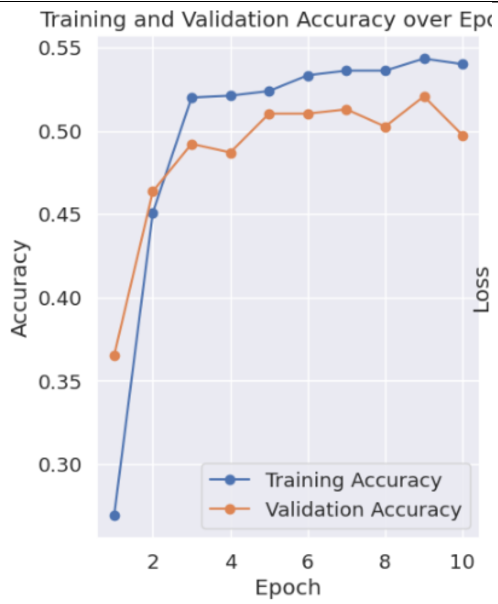
DenseNet121	 Training Accuracy: 89.21% Validation Accuracy: 81.87% Testing Accuracy: 83.16%	 Training Accuracy: 80.92% Validation Accuracy: 75.39% Testing Accuracy: 74.09%	 Training Accuracy: 74.14% Validation Accuracy: 68.91% Testing Accuracy: 63.73%

4.3.2 Learning Rate

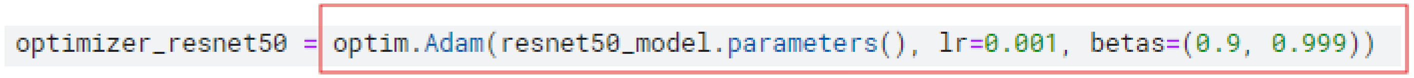
Table 4.7 Comparison for model performance with different learning rate

Architecture	Learning Rate		
	0.01	0.001	0.0001
ResNet50	 Training Accuracy: 89.38% Validation Accuracy: 80.31% Testing Accuracy: 83.94%	 Training Accuracy: 89.66% Validation Accuracy: 82.90% Testing Accuracy: 83.94%	 Training Accuracy: 53.95% Validation Accuracy: 51.55% Testing Accuracy: 47.41%

ResNet101	 Training Accuracy: 86.99% Validation Accuracy: 81.09% Testing Accuracy: 83.16%	 Training Accuracy: 90.43% Validation Accuracy: 82.90% Testing Accuracy: 84.72%	 Training Accuracy: 52.95% Validation Accuracy:54.66% Testing Accuracy: 51.55%

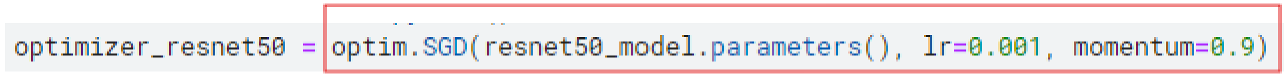
DenseNet121	 Training Accuracy: 89.21% Validation Accuracy: 78.50% Testing Accuracy: 82.38%	 Training Accuracy: 89.21% Validation Accuracy: 81.87% Testing Accuracy: 83.16%	 Training Accuracy:54.00 % Validation Accuracy: 49.74%% Testing Accuracy: 50.26%

4.3.3 Optimizer



```
optimizer_resnet50 = optim.Adam(resnet50_model.parameters(), lr=0.001, betas=(0.9, 0.999))
```

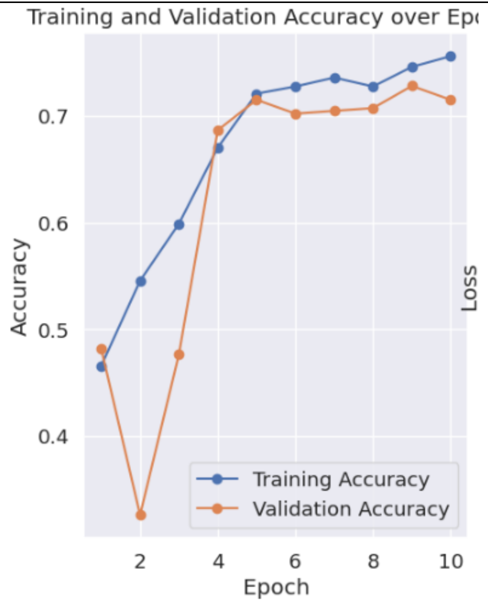
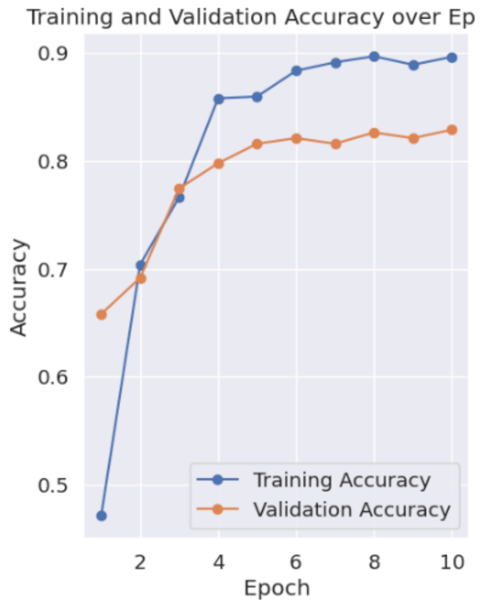
Figure 4.2.1 Code for implementation of Adam's optimizer

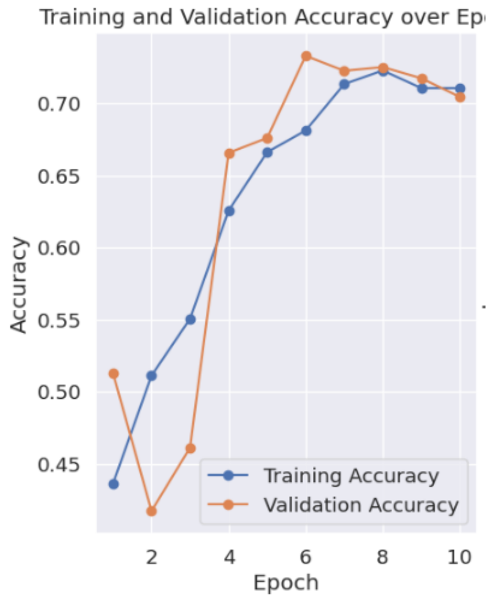
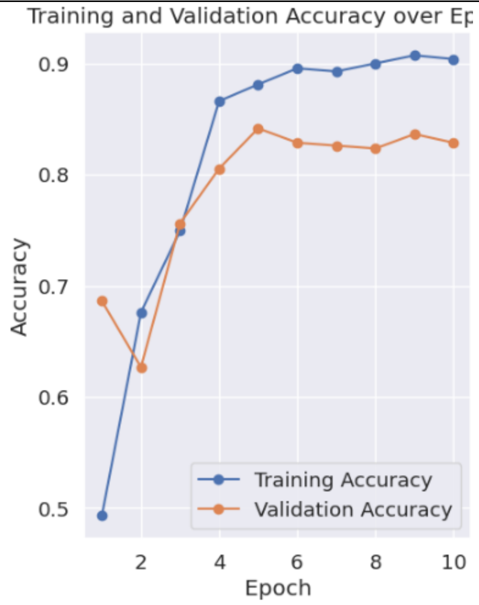


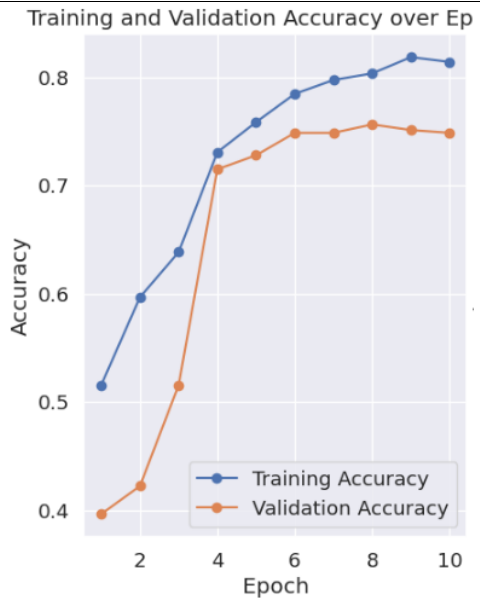
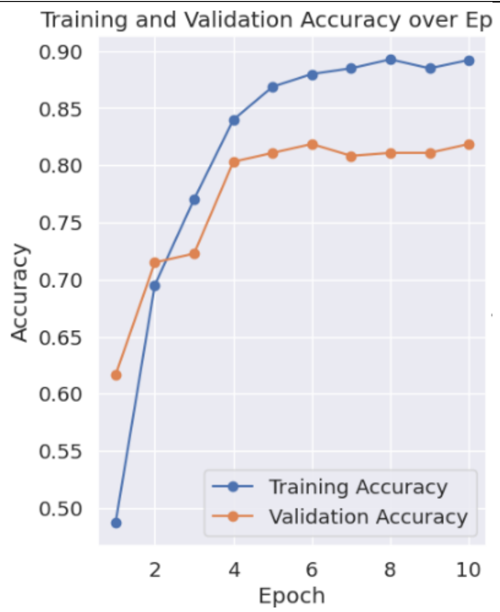
```
optimizer_resnet50 = optim.SGD(resnet50_model.parameters(), lr=0.001, momentum=0.9)
```

Figure 4.2.2 Code for implementation of SGD optimizer

Table 4.8 Comparison for model performance with different optimizer

Architecture	Optimizer	
	Adam	**Stochastic Gradient Descent (SGD)**
ResNet50	 Training Accuracy: 75.58% Validation Accuracy: 71.50% Testing Accuracy: 72.28%	 Training Accuracy: 89.66% Validation Accuracy: 82.90% Testing Accuracy: 83.94%

ResNet101	 Training Accuracy: 71.08% Validation Accuracy: 70.47% Testing Accuracy: 69.17%	 Training Accuracy: 90.43% Validation Accuracy: 82.90% Testing Accuracy: 84.72%

DenseNet121	 Training Accuracy: 81.42% Validation Accuracy: 74.87% Testing Accuracy: 82.12%	 Training Accuracy: 89.21% Validation Accuracy: 81.87% Testing Accuracy: 83.16%

After inspecting the model training process, there are several findings regarding the suitable hyperparameters for the current task. The best performance was achieved with a batch size of 16. A smaller batch size like 16 allows the model to update its weights more frequently during training, which can lead to more accurate convergence. This frequent weight update helps the model adapt quickly to complex patterns in the knee X-ray images, which might be particularly important for detecting subtle differences in osteoarthritis stages. Larger batch sizes, like 32 and 64, perform fewer weight updates, which can make the learning process less adaptive, potentially missing fine details in the data.

Learning rates of 0.01 and 0.001 performed similarly, likely because they are both within the optimal range for your model's convergence. These values allow the model to make reasonably sized updates to its weights, balancing learning speed and stability. However, after training for few rounds, it is observed that the result for learning rate 0.001 is more stable and can consistently let the model achieve >80% accuracy. On the other hand, a learning rate of 0.0001 performed much worse. This is because it is too small, resulting in very slow updates to the model’s weights, which can prevent the model from adequately learning the underlying patterns in the dataset, especially in a time-sensitive task like classification.

Stochastic Gradient Descent (SGD) outperformed Adam for all of the tested models. While Adam generally converges faster due to its adaptive learning rate, SGD often generalizes better in tasks involving smaller datasets or datasets with complex patterns like medical images. The knee osteoarthritis classification task may benefit from SGD's simplicity and more stable convergence path, leading to better generalization. Adam's faster convergence might have caused overfitting or poor generalization in this case.

Best Training Settings for Current Task:

Model = Pretrained ResNet101

Epoch = 10

Batch Size = 16

Learning Rate = 0.001

Optimizer = Stochastic Gradient Descent (SGD)

Dataset = Balanced Sharpened Dataset

Chapter 5 System Evaluation

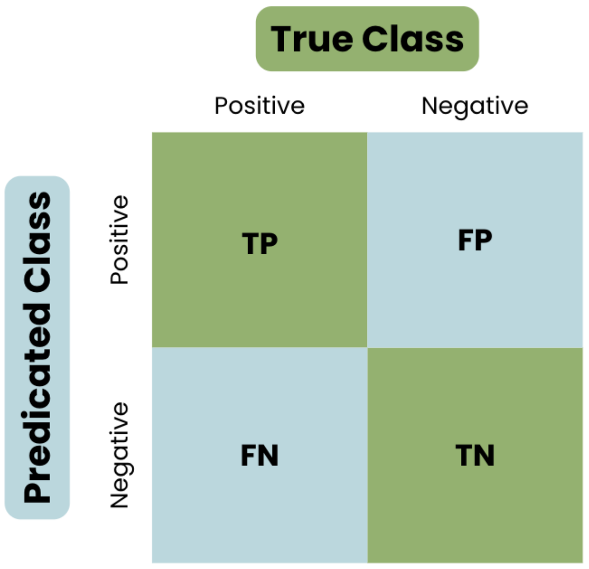
5.1. System Testing and Performance Metrics

Accuracy

$$Accuracy = \frac{True\ Positive + True\ Negative}{All\ Predictions}$$

Accuracy measures the overall correctness of a diagnostic model by comparing the number of correctly predicted knee osteoarthritis grades to the total number of cases. It is expressed as a percentage and helps gauge how often the model's predictions align with the actual grading. The grading is said to be accurate if the knee Xray image does belong to the specific class of OA.

Confusion Matrix



A confusion matrix is a table that outlines the performance of a classification model. For knee osteoarthritis grading, it displays the true positive, true negative, false positive, and false negative counts for each grade. This matrix helps visualize how well the model distinguishes between different levels of osteoarthritis severity.

Specificity

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

Specificity refers to the model's ability to correctly identify cases that do not have a particular grade of osteoarthritis. It measures the proportion of true negatives (accurately predicted non-osteoarthritis cases) out of all actual negatives. High specificity indicates fewer false positives.

Sensitivity

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Sensitivity, also known as recall, measures the model's ability to correctly identify cases with a specific grade of osteoarthritis. It is the proportion of true positives (accurately predicted osteoarthritis cases) out of all actual positives. High sensitivity means fewer false negatives.

Precision

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Precision measures the proportion of true positives out of all predicted positives. For knee osteoarthritis grading, it indicates how often the model's positive predictions (e.g., a specific grade of osteoarthritis) are correct. High precision means fewer false positives among the predicted cases.

F1 Score

$$F1\ Score = \frac{2\ \times Precision\ \times Sensitivity}{Precision + Sensitivity}$$

The F1 score is a metric that combines both precision and sensitivity into a single value, providing a balanced measure of a model's performance. It is particularly useful when dealing with imbalanced datasets, where one class (e.g., a specific grade of knee osteoarthritis) might be underrepresented. In the context of knee osteoarthritis grading, the F1 score helps evaluate how well the model performs in correctly identifying both the presence and absence of different grades of osteoarthritis. A higher F1 score indicates a better balance between precision and sensitivity, meaning the model is good at both avoiding false positives and identifying true positives.

Cross Entropy Loss

$$H(p,q) = \Sigma\, p(x) \log q(x)$$

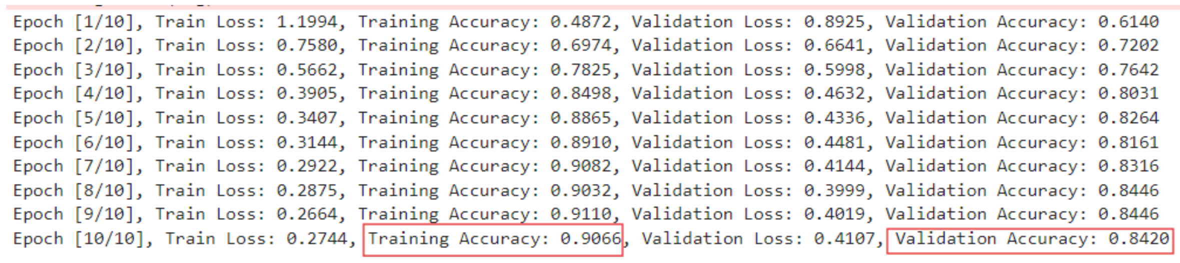
Cross entropy loss quantifies the difference between the predicted probability distribution and the actual distribution (true labels) in classification tasks. In knee osteoarthritis grading, it assesses how well the model's predicted probabilities match the actual grades, with lower loss indicating better model performance.

5.2 Final Testing Setup and Results

Final System Setup

Model = Pretrained ResNet101
Dataset = Balanced Sharpened Dataset
Epoch = 10
Batch Size = 16
Learning Rate = 0.001
Optimizer = Stochastic Gradient Descent (SGD)

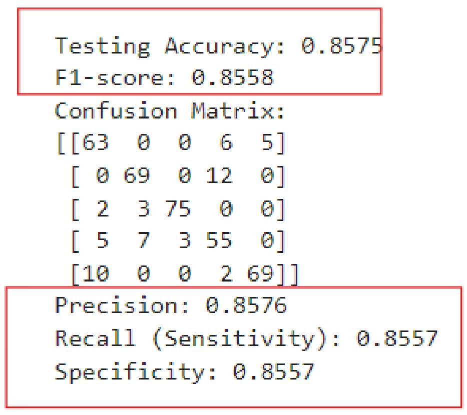
Final Training & Testing Result



```
Epoch [1/10], Train Loss: 1.1994, Training Accuracy: 0.4872, Validation Loss: 0.8925, Validation Accuracy: 0.6140
Epoch [2/10], Train Loss: 0.7580, Training Accuracy: 0.6974, Validation Loss: 0.6641, Validation Accuracy: 0.7202
Epoch [3/10], Train Loss: 0.5662, Training Accuracy: 0.7825, Validation Loss: 0.5998, Validation Accuracy: 0.7642
Epoch [4/10], Train Loss: 0.3905, Training Accuracy: 0.8498, Validation Loss: 0.4632, Validation Accuracy: 0.8031
Epoch [5/10], Train Loss: 0.3407, Training Accuracy: 0.8865, Validation Loss: 0.4336, Validation Accuracy: 0.8264
Epoch [6/10], Train Loss: 0.3144, Training Accuracy: 0.8910, Validation Loss: 0.4481, Validation Accuracy: 0.8161
Epoch [7/10], Train Loss: 0.2922, Training Accuracy: 0.9082, Validation Loss: 0.4144, Validation Accuracy: 0.8316
Epoch [8/10], Train Loss: 0.2875, Training Accuracy: 0.9032, Validation Loss: 0.3999, Validation Accuracy: 0.8446
Epoch [9/10], Train Loss: 0.2664, Training Accuracy: 0.9110, Validation Loss: 0.4019, Validation Accuracy: 0.8446
Epoch [10/10], Train Loss: 0.2744, Training Accuracy: 0.9066, Validation Loss: 0.4107, Validation Accuracy: 0.8420
```

Figure 5.2.1 Training history for final model

Over 10 epochs, the fine-tuned pretrained ResNet101 model shows significant improvement, with training loss decreasing from 1.1994 to 0.2744 and training accuracy rising from 48.72% to 90.66%. Validation loss also decreases from 0.8925 to 0.4107, and validation accuracy improves from 61.40% to 84.20%. Despite these gains, fluctuations in validation loss suggest early signs of overfitting, where the model fits training data well but faces challenges in generalizing. By the final epoch, the model exhibits strong performance with high accuracy and stabilized validation loss, indicating effective learning.



```
Testing Accuracy: 0.8575
F1-score: 0.8558
Confusion Matrix:
[[63  0  0  6  5]
 [ 0 69  0 12  0]
 [ 2  3 75  0  0]
 [ 5  7  3 55  0]
 [10  0  0  2 69]]
Precision: 0.8576
Recall (Sensitivity): 0.8557
Specificity: 0.8557
```

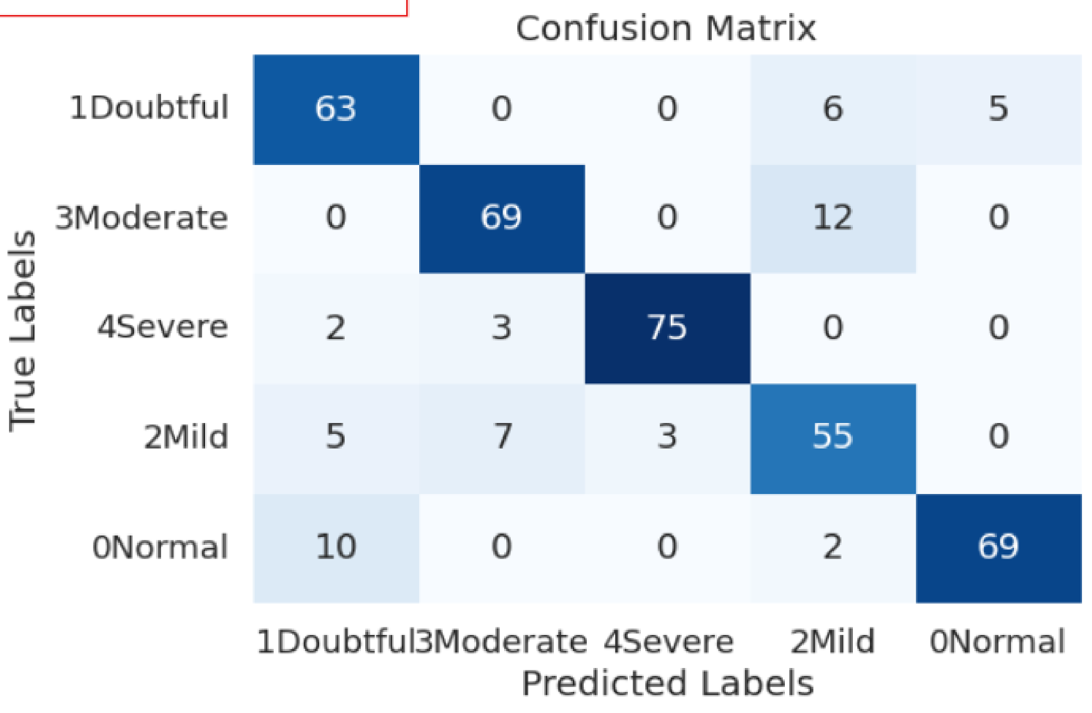
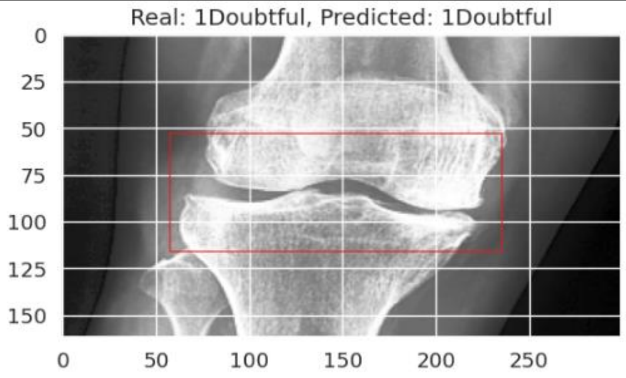
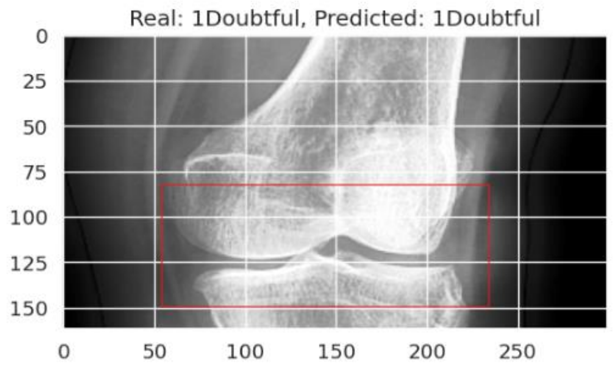
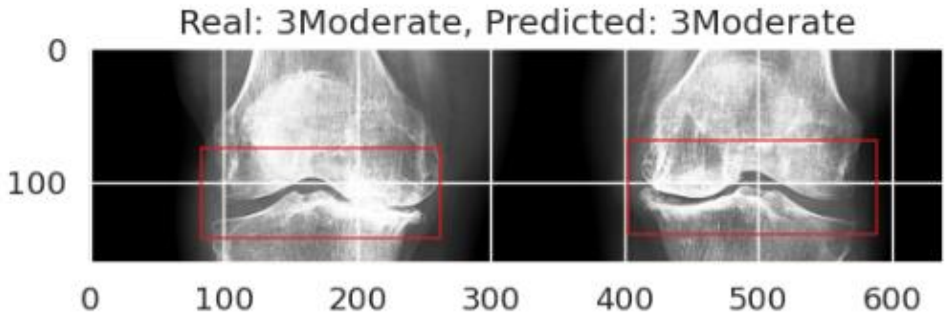
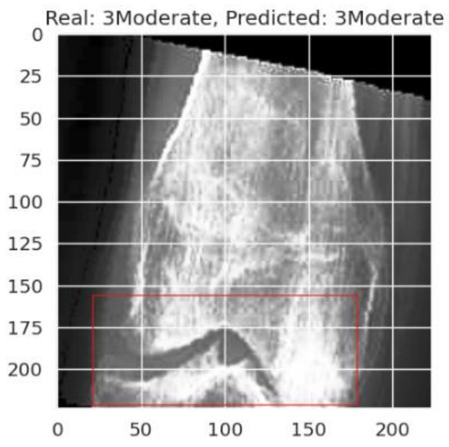


Figure 5.2.2 Final testing result for final model

The final testing results for the model show a testing accuracy of 85.75% and an F1-score of 0.8558, indicating strong overall performance. The confusion matrix reveals that the model performs exceptionally well in distinguishing between classes, with high precision (0.8576) and recall (0.8557) across the categories. The specificity, also at 0.8557, reflects a balanced ability to identify true negatives across the classes. The model shows robust classification performance with effective detection and differentiation of various classes, though there is room for further refinement to address any misclassifications and enhance overall precision and recall.

Example of Knee Osteoarthritis Grading Result

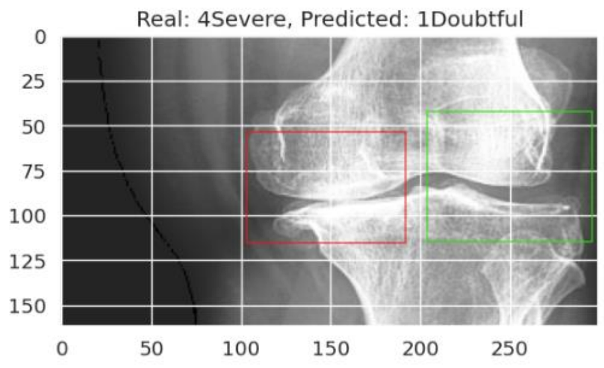
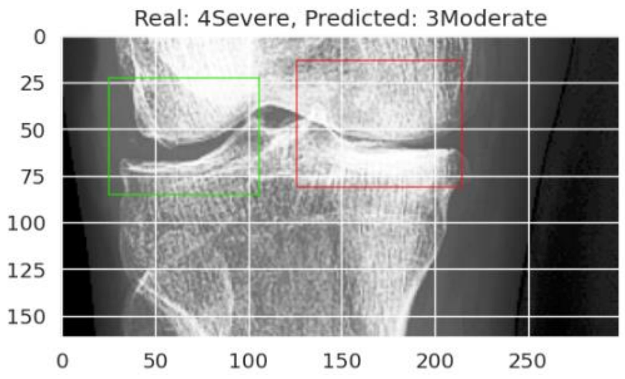
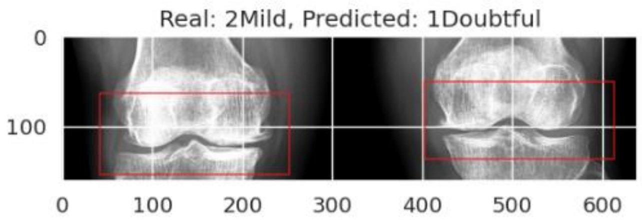
Table 5.1 Example of knee OA grading

1.	
2.	
3.	
4.	

5.3 Error analysis

The wrongly predicted classes were printed to observe any potential issue that cause the incorrect predictions.

Table 5.2 Error analysis for misclassified classes

	Prediction Output	Possible Explanation
1		The model extracted feature from the green bounding box where the gap between bones are still large instead of red bounding box where the bones touch each other causing wrong prediction.
2		The difference between grade 3 and 4 OA is too small for the model to identify.
3		The presence of 2 knee joint make the model confused about which side it should grade.

5.4 Objective Evaluation

The research objectives have been effectively achieved, demonstrating significant progress in grading the severity of knee osteoarthritis (OA) from X-ray images using deep learning techniques. The comparative analysis of various models highlighted that the pretrained ResNet101 emerged as the best-performing classifier for this task, as evidenced by its testing accuracy of 85.75% and an F1-score of 0.8558. The confusion matrix reveals that the model consistently distinguishes between different OA severity levels, with high precision (0.8576), recall (0.8557), and specificity (0.8557). This confirms that ResNet101 excels in classifying the severity of knee OA according to the KL grading system, offering robust performance in automated grading.

Additionally, the optimization efforts, including architecture selection, hyperparameter tuning, and data augmentation, contributed significantly to enhancing the model's accuracy and sensitivity. The implementation of transfer learning further improved model performance by accelerating convergence and enhancing generalization, reducing data requirements, and providing better feature representation. These outcomes validate the effectiveness of the employed strategies and confirm that the research objectives were successfully met, positioning the pretrained ResNet101 model as a reliable tool for automated OA grading from X-ray images.

Chapter 6 Conclusion and Future Work

6.1 Conclusion

For conclusion, all of the objective stated from the start of project has been fulfilled. The pretrained ResNet101 perform the best with final accuracy of 85.75% without any overfitting issues. The model also achieved high F1 score (0.8558), precision (0.8576), recall (0.8557) and specificity (0.8557). Compared with the work of Aslan [9] with same dataset that uses deep learning model for feature extraction, NCA for feature selection and finally machine learning models for classification, this paper chose to improve the dataset at the first place and uses only pretrained model for the knee OA grading task. The achieved accuracy of 85.75% is slightly higher compared to 84.12% accuracy in Aslan [9] work mostly due to the improvement done in dataset. It can be seen that without building up bulky model and passing the data through few sections, using only on deep learning model can also achieve similar accuracy or even higher. This also proved the importance of dataset quality in model training. Although accuracy of 85.75% is still far from perfect, but the usage of imperfect dataset with only 1650 images were only used to achieve this performance while those in the reviewed articles are mostly trained using OAI dataset (~45000 x-ray images) and MOST dataset (~14000 x-ray images). Data augmentation has been used to balance out the initially slanted dataset and the result of model training leap up a big stage after so.

In terms of AUC for each grade of OA, the model has achieved 0.92, 0.90, 0.86, 0.91, 0.96 AUC for each grade of OA from 0 to 4 respectively. The improvement in classifying grade 2 OA improved the most from 68.10% in Aslan [9] work to 78.57% in current work. This improvement is important as OA is incurable and only early detection can bring the most benefit.

Thirteen different models were trained for the knee OA grading task to determine which performs best. These models include EfficientNet-B0, B3, EfficientNetV2, ResNet50, ResNet101, DenseNet121, DenseNet161, VGG-19, VGG-16, GoogLeNet, SqueezeNet, RegNet, and AlexNet. Each model has distinct architecture and performance characteristics. EfficientNet models are known for their efficiency and scalability, balancing accuracy with computational needs. ResNet models use deep residual learning for better performance with fewer layers, while DenseNet models improve feature reuse at a higher computational cost. VGG models, though deep and effective, are resource-heavy due to many parameters. GoogLeNet, SqueezeNet, and RegNet offer trade-offs between complexity and performance, with AlexNet serving as a simpler baseline model. Each model's strengths and weaknesses affect their knee OA grading performance differently.

After that, fine-tuning of model was done on batch size, learning rate and optimizer used. The performance of model that varies based on different hyperparameter also shown the importance of using the optimal hyperparameters. All of the steps mentioned has led the model training to a successful end.

6.2 Recommendation

Based on the findings of this research, it is recommended that future work continues to explore deep learning models for knee osteoarthritis (OA) grading, particularly leveraging transfer learning techniques to further enhance model performance. The pretrained ResNet101 model that achieved accuracy of 85.75%, slightly higher (~1.5%) compared to DenseNet201 in Aslan’s work [9] achieving 84.12% accuracy, but with a more straightforward method, using only pretrained and fine-tuned ResNet101, without the extended feature engineering. This demonstrated strong performance in both accuracy and generalization, making the model a viable candidate for clinical applications. However, incorporating additional optimization strategies, such as advanced data augmentation, fine-tuning hyperparameters, and using ensemble methods, could further improve the model's ability to detect subtle OA features. Exploring other emerging deep learning architectures, such as Vision Transformers (ViTs), could also provide new avenues for improving performance.

Moreover, to increase the model's robustness and applicability in real-world settings, it is recommended to use larger and more diverse datasets that represent a wide range of OA severities and patient demographics for example OAI and MOST. This would help address potential biases and improve the model’s generalization to broader populations. Additionally, integrating multimodal data such as patient history and clinical data, alongside X-ray images, could provide more context and enhance the predictive accuracy of the models. Future studies should also focus on evaluating these models in clinical settings to ensure their reliability, interpretability, and effectiveness for aiding in the diagnosis and treatment of knee OA.

REFERENCES

[1] V. L. Johnson and D. J. Hunter, “The epidemiology of osteoarthritis,” Best Practice & Research Clinical Rheumatology, vol. 28, no. 1, pp. 5–15, Feb. 2014, doi: https://doi.org/10.1016/j.berh.2014.01.004.

[2] S. A. Ali, K. E. Walsh, and M. Kloseck, “Patient perspectives on improving osteoarthritis management in urban and rural communities,” Journal of Pain Research, vol. Volume 11, pp. 417–425, Feb. 2018, doi: https://doi.org/10.2147/jpr.s150578.

[3] F. Jacobs, “History of Radiology: Timeline, Pioneers, Inventions | RamSoft,” RamSoft Inc. Healthcare IT, Dec. 13, 2021. https://www.ramsoft.com/history-of-radiology

[4] A. J. Schwartz, H. D. Clarke, M. J. Spangehl, J. S. Bingham, D. A. Etzioni, and M. R. Neville, “Can a Convolutional Neural Network Classify Knee Osteoarthritis on Plain Radiographs as Accurately as Fellowship-Trained Knee Arthroplasty Surgeons?,” The Journal of Arthroplasty, vol. 35, no. 9, pp. 2423–2428, Sep. 2020, doi: https://doi.org/10.1016/j.arth.2020.04.059.

[5] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala, “Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach,” Scientific Reports, vol. 8, no. 1, Jan. 2018, doi: https://doi.org/10.1038/s41598-018-20132-7.

[6] H. H. Nguyen, S. Saarakkala, M. B. Blaschko, and A. Tiulpin, “Semixup: In- and Out-of-Manifold Regularization for Deep Semi-Supervised Knee Osteoarthritis Severity Grading From Plain Radiographs,” IEEE Transactions on Medical Imaging, vol. 39, no. 12, pp. 4346–4356, Dec. 2020, doi: https://doi.org/10.1109/TMI.2020.3017007.

[7] B. Norman, V. Pedoia, A. Noworolski, T. M. Link, and S. Majumdar, “Applying Densely Connected Convolutional Neural Networks for Staging Osteoarthritis Severity from Plain Radiographs,” Journal of Digital Imaging, vol. 32, no. 3, pp. 471–477, Oct. 2018, doi: https://doi.org/10.1007/s10278-018-0098-3.

[8] V. Pedoia, J. Lee, B. Norman, T. M. Link, and S. Majumdar, “Diagnosing osteoarthritis from T2 maps using deep learning: an analysis of the entire Osteoarthritis Initiative baseline cohort,” Osteoarthritis and Cartilage, vol. 27, no. 7, pp. 1002–1010, Jul. 2019, doi: https://doi.org/10.1016/j.joca.2019.02.800.

[9] S. Aslan, “Automatic Detection of Knee Osteoarthritis Disease with the Developed CNN, NCA and SVM Based Hybrid Model,” Traitement du Signal, vol. 40, no. 1, pp. 317–326, Feb. 2023, doi: https://doi.org/10.18280/ts.400131.

[10] S. Olsson, E. Akbarian, A. Lind, A. S. Razavian, and M. Gordon, “Automating classification of osteoarthritis according to Kellgren-Lawrence in the knee using deep learning in an unfiltered adult population,” BMC Musculoskeletal Disorders, vol. 22, no. 1, Oct. 2021, doi: https://doi.org/10.1186/s12891-021-04722-7.

[11] D. H. Kim, K. J. Lee, D. Choi, J. I. Lee, H. G. Choi, and Y. S. Lee, “Can Additional Patient Information Improve the Diagnostic Performance of Deep Learning for the Interpretation of Knee Osteoarthritis Severity,” Journal of Clinical Medicine, vol. 9, no. 10, p. 3341, Oct. 2020, doi: https://doi.org/10.3390/jcm9103341.

[12] A. Tiulpin and S. Saarakkala, “Automatic Grading of Individual Knee Osteoarthritis Features in Plain Radiographs Using Deep Convolutional Neural Networks,” Diagnostics, vol. 10, no. 11, p. 932, Nov. 2020, doi: https://doi.org/10.3390/diagnostics10110932.

[13] M. Górriz, J. Antony, K. McGuinness, X. Giró-i-Nieto, and N. E. O’Connor, “Assessing Knee OA Severity with CNN attention-based end-to-end architectures,” proceedings.mlr.press, May 24, 2019. https://proceedings.mlr.press/v102/gorriz19a.html

[14] J. Antony, K. McGuinness, N. E. O’Connor, and K. Moran, “Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks,” 2016 23rd International Conference on Pattern Recognition (ICPR), Dec. 2016, doi: https://doi.org/10.1109/icpr.2016.7899799.

[15] P. Chen, L. Gao, X. Shi, K. Allen, and L. Yang, "Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss," Computerized Medical Imaging and Graphics, vol. 75, pp. 84–92, Jul. 2019, doi: https://doi.org/10.1016/j.compmedimag.2019.06.002.

[16] "Digital Knee X-ray," www.kaggle.com. https://www.kaggle.com/datasets/tommyngx/digital-knee-xray

APPENDIX A

A1 Weekly Report

FINAL YEAR PROJECT WEEKLY REPORT
(Project II)

Trimester, Year: TRIMESTER 2 YEAR 3	**Study week no.: 2**
Student Name & ID: Jevyline Ng (21ACB00637)	
Supervisor: Dr. Ts. Goh Chuan Meng	
Project Title: Knee Osteoarthritis Grading Using Deep Learning Classifier	

1. WORK DONE

- Revise back the work done on FYP1
- Try to find dataset used from any of the reviewed articles

2. WORK TO BE DONE

- Choose suitable model to train
- Work on the dataset found, inspecting, cleaning
- Perform Data pre-processing

3. PROBLEMS ENCOUNTERED

- Tried to manually code layers of the model but failed
- Good dataset used by reviewed articles require permission to access

4. SELF EVALUATION OF THE PROGRESS

- Did not do anything during semester break. Procrastination causes more work to be done after class.

[illegible] [illegible]

Supervisor's signature Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: TRIMESTER 2 YEAR 3	**Study week no.: 4**
Student Name & ID: Jevyline Ng (21ACB00637)	
Supervisor: Dr. Ts. Goh Chuan Meng	
Project Title: Knee Osteoarthritis Grading Using Deep Learning Classifier	

1. WORK DONE

- Found a dataset from Kaggle that can be used for model training
- Setup data preprocessing pipeline

2. WORK TO BE DONE

- Select several deep learning models that are pretrained
- Set up training environment in Kaggle using Pytorch/ Tensorflow

3. PROBLEMS ENCOUNTERED

- Lack knowledge about available model
- Lack knowledge about model structure and how to implement

4. SELF EVALUATION OF THE PROGRESS

- Try to learn how to use PyTorch as fast as possible to start coding work. Quite good.

[illegible]

Supervisor's signature

[illegible]

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: TRIMESTER 2 YEAR 3	**Study week no.: 6**
Student Name & ID: Jevyline Ng (21ACB00637)	
Supervisor: Dr. Ts. Goh Chuan Meng	
Project Title: Knee Osteoarthritis Grading Using Deep Learning Classifier	

1. WORK DONE

- Trained few models such as ResNet, VGG and AlexNet
- Successfully setup training environment in Kaggle using Pytorch

2. WORK TO BE DONE

- Train more model
- Improve model performance
- Improve dataset

3. PROBLEMS ENCOUNTERED

- Found out model performance not good, accuracy low
- Found out dataset quality is not high, causing training stagnant

4. SELF EVALUATION OF THE PROGRESS

- Tried to learn more model as fast as possible. Quite good

[illegible]
Supervisor's signature

[illegible]
Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: TRIMESTER 2 YEAR 3	**Study week no.: 8**
Student Name & ID: Jevyline Ng (21ACB00637)	
Supervisor: Dr. Ts. Goh Chuan Meng	
Project Title: Knee Osteoarthritis Grading Using Deep Learning Classifier	

1. WORK DONE

- Train more kind of model, more variants and more new models
- Implemented transfer learning
- Solved data imbalance problem

2. WORK TO BE DONE

- Perform model fine tuning to find out the best one
- Maybe include more model since current performance is not so good

3. PROBLEMS ENCOUNTERED

- Model performance stay stagnant, accuracy not high enough
- Hard to find any method to breakthrough current performance

4. SELF EVALUATION OF THE PROGRESS

- Figured out way to improve the model with trying to improve dataset, obtained quite good result. Otherwise, didn't do much due to procrastination.

[illegible] [illegible]

Supervisor's signature   Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: TRIMESTER 2 YEAR 3	**Study week no.: 10**
Student Name & ID: Jevyline Ng (21ACB00637)	
Supervisor: Dr. Ts. Goh Chuan Meng	
Project Title: Knee Osteoarthritis Grading Using Deep Learning Classifier	

1. WORK DONE

- Done fine tuning the models
- Included more model in training to have more options

2. WORK TO BE DONE

- Finalizing the research
- Report writing

3. PROBLEMS ENCOUNTERED

- Fine-tuning takes up a lot of time
- Need to run the model for very long time

4. SELF EVALUATION OF THE PROGRESS

- Spent a lot of time tuning the model as running each time need quite some time. Quite good

[illegible]

Supervisor's signature

[illegible]

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: TRIMESTER 2 YEAR 3	**Study week no.: 13**
Student Name & ID: Jevyline Ng (21ACB00637)	
Supervisor: Dr. Ts. Goh Chuan Meng	
Project Title: Knee Osteoarthritis Grading Using Deep Learning Classifier	

1. WORK DONE

- Done finalizing model
- Started writing report

2. WORK TO BE DONE

- Finish Report
- Prepare for presentation

3. PROBLEMS ENCOUNTERED

- Slow progress due to lack of time doing other assignments and presentation

4. SELF EVALUATION OF THE PROGRESS

- Tried my best to complete project within time given, quite good

[illegible]

Supervisor's signature

[illegible]

Student's signature

A2 Poster

Knee Osteoarthritis Grading Using Deep Learning Classifier

Name: Jevyline Ng   Student ID: 21ACB00637   Faculty/Programme: FICT/CS

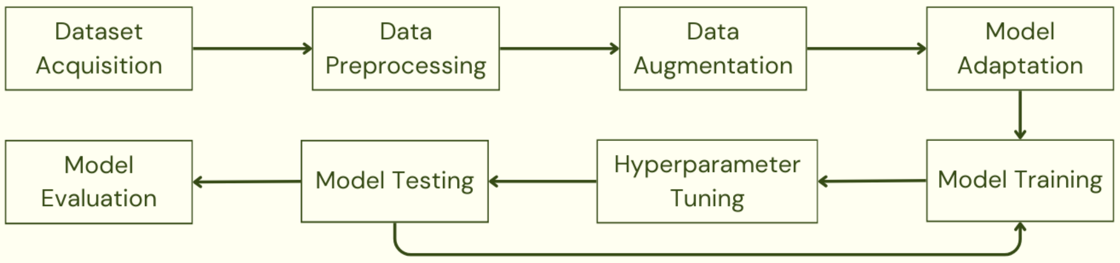
Introduction

Osteoarthritis (OA) is a prevalent joint disease affecting various populations, not just the elderly, with factors like genetics, obesity, and injuries contributing to its incidence. **While incurable, early detection can slow its progression. Artificial Intelligence (AI) offers promise for early OA detection**, but further research is needed to **optimize AI models for diverse patient populations.** Hence, the research aims to **find the best model** to perform the knee OA grading task.

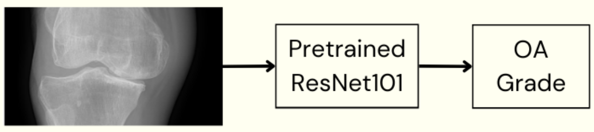
Objectives

- Find the best deep learning classifier to grade severity of OA on the knee X-ray images.
- Optimize the deep learning model to enhance its accuracy and sensitivity in identifying OA features on X-ray images.
- Implementing transfer learning to leverage the model overall performance

Methodology



Result & Discussion



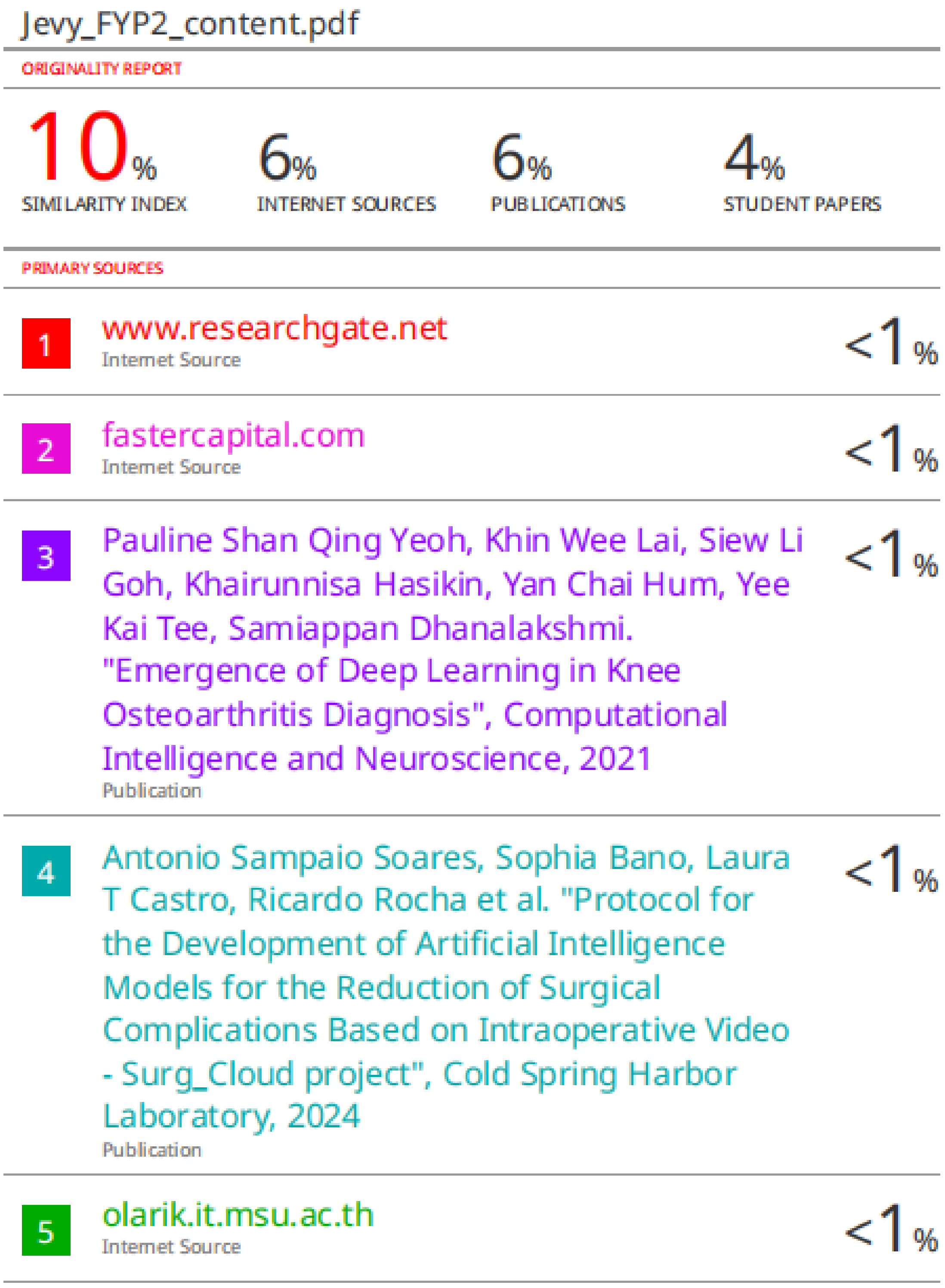
Accuracy: 0.8290   **Precision: 0.8292**
Sensitivity: 0.8273   **Specificity: 0.8273**

The model shows strong performance, with solid accuracy and an F1-score reflecting good overall classification. It handles class distinctions well, with high precision, recall, and specificity. While the results are promising, there is still room for improvement to reduce misclassifications and further boost precision and recall.

Conclusion

- The project successfully met all objectives, with ResNet101 performing best, achieving high accuracy and metrics without overfitting, emphasizing the importance of dataset quality and augmentation over complex models.
- 13 pretrained models were tested, and fine-tuning hyperparameters like batch size and learning rate played a crucial role in optimizing performance.
- Future work should explore transfer learning, advanced optimization techniques, larger diverse datasets, and multimodal data integration to further enhance model performance and ensure clinical applicability in knee OA grading.

Plagiarism Check Result



Jevy_FYP2_content.pdf

ORIGINALITY REPORT

10%	6%	6%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchgate.net Internet Source	<1%
2	fastercapital.com Internet Source	<1%
3	Pauline Shan Qing Yeoh, Khin Wee Lai, Siew Li Goh, Khairunnisa Hasikin, Yan Chai Hum, Yee Kai Tee, Samiappan Dhanalakshmi. "Emergence of Deep Learning in Knee Osteoarthritis Diagnosis", Computational Intelligence and Neuroscience, 2021 Publication	<1%
4	Antonio Sampaio Soares, Sophia Bano, Laura T Castro, Ricardo Rocha et al. "Protocol for the Development of Artificial Intelligence Models for the Reduction of Surgical Complications Based on Intraoperative Video - Surg_Cloud project", Cold Spring Harbor Laboratory, 2024 Publication	<1%
5	olarik.it.msu.ac.th Internet Source	<1%

Universiti Tunku Abdul Rahman			
Form Title : Supervisor's Comments on Originality Report Generated by Turnitin for Submission of Final Year Project Report (for Undergraduate Programmes)			
Form Number: FM-IAD-005	Rev No.: 0	Effective Date: 01/10/2013	Page No.: 1of 1

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Full Name(s) of Candidate(s)	Jevyline Ng
ID Number(s)	21ACB00637
Programme / Course	Bachelor of Computer Science (Honours)
Title of Final Year Project	Knee Osteoarthritis Grading Using Deep Learning Classifier

Similarity	**Supervisor's Comments (Compulsory if parameters of originality exceeds the limits approved by UTAR)**
Overall similarity index: 10 % **Similarity by source** Internet Sources: 6 % Publications: 6 % Student Papers: 4 %	
Number of individual sources listed of more than 3% similarity: 0	
Parameters of originality required and limits approved by UTAR are as Follows: **(i) Overall similarity index is 20% and below, and** **(ii) Matching of individual sources listed must be less than 3% each, and** **(iii) Matching texts in continuous block must not exceed 8 words** ***Note: Parameters (i) – (ii) shall exclude quotes, bibliography and text matches which are less than 8 words.***	

Note Supervisor/Candidate(s) is/are required to provide softcopy of full set of the originality report to Faculty/Institute

Based on the above results, I hereby declare that I am satisfied with the originality of the Final Year Project Report submitted by my student(s) as named above.

[illegible]	
Signature of Supervisor	Signature of Co-Supervisor
Name: Ts. Dr. Goh Chuan Meng	Name:
Date: 13/09/2024	Date:

UNIVERSITI TUNKU ABDUL RAHMAN

FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY (KAMPAR CAMPUS)

CHECKLIST FOR FYP2 THESIS SUBMISSION

Student Id	21ACB00637
Student Name	JEVYLINE NG
Supervisor Name	TS. DR. GOH CHUAN MENG

TICK (√)	**DOCUMENT ITEMS** Your report must include all the items below. Put a tick on the left column after you have checked your report with respect to the corresponding item.
√	Title Page
√	Signed Report Status Declaration Form
√	Signed FYP Thesis Submission Form
√	Signed form of the Declaration of Originality
√	Acknowledgement
√	Abstract
√	Table of Contents
√	List of Figures (if applicable)
√	List of Tables (if applicable)
	List of Symbols (if applicable)
√	List of Abbreviations (if applicable)
√	Chapters / Content
√	Bibliography (or References)
√	All references in bibliography are cited in the thesis, especially in the chapter of literature review
√	Appendices (if applicable)
√	Weekly Log
√	Poster
√	Signed Turnitin Report (Plagiarism Check Result - Form Number: FM-IAD-005)
√	I agree 5 marks will be deducted due to incorrect format, declare wrongly the ticked of these items, and/or any dispute happening for these items in this report.

*Include this form (checklist) in the thesis (Bind together as the last page)

I, the author, have checked and confirmed all the items listed in the table are included in my report.

[illegible]
____
(Signature of Student)
Date: 13/09/2024