AI-based Knee Osteoarthritis Severity Classification

Table of Contents

Table of Figures

Abstract

Knee osteoarthritis (OA) is a widespread musculoskeletal disorder that predominantly affects older adults, leading to chronic pain, reduced mobility, and significant socioeconomic burden (Hunter & Bierma-Zeinstra, 2019). Accurate and early diagnosis is critical for initiating effective treatments; however, conventional radiographic assessments often suffer from subjectivity and diagnostic inconsistencies among clinicians (Katz et al., 2021). To address this, the current research explores the application of deep learning techniques in automating the classification of knee OA severity using X-ray imaging.

This study proposes a deep learning-based classification model, developed using transfer learning with two advanced convolutional neural network (CNN) architectures—EfficientNetB3 and ResNet50. The models were fine-tuned on a publicly available Kaggle dataset, consisting of over 7,000 labelled knee X-ray images. Extensive preprocessing steps, such as data augmentation, grayscale normalization, and batch normalization, were incorporated to enhance model generalizability and reduce overfitting. Additionally, interpretability was supported through the integration of Grad-CAM visualizations to help clinicians understand the network's focus regions (Selvaraju et al., 2017).

Experimental results revealed that EfficientNetB3 achieved a test accuracy of 97%, outperforming ResNet50, which achieved 87%. Precision, recall, and F1-scores across all classes further validated the reliability of both models. These findings highlight the practical potential of deploying AI-assisted tools in radiological workflows, especially in resource-constrained settings.

Importantly, the research adheres to ethical, legal, and professional standards, ensuring that the model development process respects patient data privacy and aligns with responsible AI guidelines (Floridi et al., 2018). While the model is not a substitute for clinical diagnosis, it serves as a supportive tool aimed at enhancing diagnostic consistency and reducing workload.

In future work, further model improvements and the inclusion of larger, demographically diverse datasets are proposed to strengthen clinical applicability and fairness across patient groups.

Chapter 1 – Introduction.

Knee osteoarthritis (KOA) is a chronic and progressive joint disorder that significantly impairs mobility and reduces the quality of life for millions globally. Characterised by the gradual degradation of articular cartilage and structural alterations in the knee joint, KOA often results in persistent pain, stiffness, swelling, and functional disability. According to Hunter and Bierma-Zeinstra (2019), over 500 million people worldwide are affected by KOA, making it one of the most prevalent musculoskeletal conditions and a leading cause of disability, particularly among ageing populations.

Timely and precise diagnosis is crucial for effective disease management and to halt or slow KOA progression. Radiographic imaging, particularly knee X-rays interpreted using the Kellgren-Lawrence (KL) grading system, remains the clinical gold standard for evaluating KOA severity. However, manual grading is subject to inter-observer variability, diagnostic delay, and inconsistencies, especially in high-volume clinical environments (Kohn et al., 2016). To address these challenges, artificial intelligence (AI) and deep learning technologies have gained traction as tools capable of automating and standardising diagnostic processes in medical imaging.

Convolutional Neural Networks (CNNs) have revolutionised medical image analysis by enabling automatic feature extraction and robust classification. Advanced CNN architectures such as ResNet50 (He et al., 2016) and EfficientNetB3 (Tan & Le, 2019) have demonstrated remarkable success across various healthcare applications, including dermatological disease recognition, pulmonary condition detection, and retinal anomaly diagnosis (Esteva et al., 2017; Rajpurkar et al., 2017; Gulshan et al., 2016). Building upon this foundation, the current research investigates the efficacy of ResNet50 and EfficientNetB3 in classifying KOA severity levels using knee X-ray images from publicly available datasets.

The central objective of this project is to construct an end-to-end deep learning classification framework that accurately grades KOA severity, thereby assisting in early detection and treatment planning. Beyond accuracy, the study prioritises interpretability through the integration of Gradient-weighted Class Activation Mapping (Grad-CAM), which visualises salient regions of input images that influence model predictions. This helps bridge the gap between clinical trust and AI automation by making the decision-making process more transparent.

To enhance accessibility and usability, the final model has been deployed as an interactive web application using the Hugging Face Spaces platform. Built with Gradio, this intuitive interface allows clinicians and researchers to upload knee X-rays, receive severity predictions in real time, and view corresponding Grad-CAM heatmaps that highlight the most diagnostically relevant areas of the image. This approach supports rapid, explainable, and user-friendly KOA assessment—particularly valuable in under-resourced settings lacking specialised radiological expertise.

In addition to technical implementation, the study addresses ethical considerations such as patient data privacy, algorithmic fairness, and the importance of explainability in clinical contexts. By combining high-performance AI modelling with responsible deployment practices, this project contributes a scalable, interpretable, and practical solution for KOA detection and monitoring in real-world healthcare environments.

1.1 Background

Knee osteoarthritis (KOA) is a prevalent degenerative joint disorder and a leading cause of chronic pain and mobility impairment, particularly among the elderly. The diagnostic process for KOA traditionally begins with a clinical evaluation, followed by radiographic imaging, which remains a cornerstone for assessing the structural changes in the knee joint. One of the most widely used grading systems for evaluating KOA severity is the Kellgren-Lawrence (KL) scale, which classifies disease progression into five grades—from Grade 0 (no radiographic features) to Grade 4 (severe joint space narrowing and bone deformity) (Kellgren & Lawrence, 1957). Despite its global recognition, the KL system relies on the subjective interpretation of radiographs by trained specialists, which introduces inter-observer variability and may lead to inconsistent diagnoses or delayed treatment decisions.

The emergence of artificial intelligence (AI) and deep learning technologies offers a promising solution to this challenge. The growing availability of labelled medical image datasets, along with advancements in GPU computing, has enabled the use of Convolutional Neural Networks (CNNs) in various diagnostic imaging tasks. CNNs are particularly powerful for extracting hierarchical features from visual data, making them ideal for analysing X-ray images where subtle texture, shape, and intensity patterns correspond to different KOA stages (LeCun, Bengio & Hinton, 2015).

Two prominent CNN architectures, ResNet50 and EfficientNetB3, have shown considerable potential in the medical domain. ResNet50, introduced by He et al. (2016), leverages residual learning to allow training of deeper networks by mitigating vanishing gradient problems. This enables it to learn nuanced features necessary for distinguishing between close grades in medical images. EfficientNetB3, on the other hand, is part of the EfficientNet family designed by Tan and Le (2019). It employs a compound scaling technique that optimally balances network depth, width, and resolution. As a result, it achieves competitive accuracy while maintaining computational efficiency, making it particularly attractive for real-time or resource-constrained clinical settings.

Despite their proven capabilities, limited research has directly compared ResNet50 and EfficientNetB3 specifically for KOA severity classification. This project seeks to address this gap by implementing both models on the same dataset and evaluating them across consistent metrics, including accuracy, F1-score, and model interpretability. Importantly, this study enhances explainability using Gradient-weighted Class Activation Mapping (Grad-CAM), which provides heatmaps highlighting image regions most influential to the model's decision. Such visual interpretations are vital for increasing clinician trust and aligning AI outputs with clinical reasoning.

To ensure the practical utility of the proposed solution, the final model is deployed as a web application using Hugging Face Spaces, built with Gradio. This intuitive interface allows users—clinicians, researchers, or healthcare stakeholders—to upload knee X-rays, receive KOA severity predictions, and instantly view Grad-CAM visualisations, all within a browser-based environment. The deployment underscores the project's commitment to not only research rigor but also real-world accessibility and usability.

Through this dual-model comparative study and its web-based implementation, the research aims to identify the optimal balance between accuracy, speed, and transparency, which are essential for integrating AI-assisted diagnosis into everyday healthcare workflows—especially in under-resourced or high-throughput environments.

1.2 Problem Statement

Knee osteoarthritis (KOA) is among the most widespread musculoskeletal disorders globally, exerting a profound impact on the quality of life of the ageing population. Marked by the progressive deterioration of articular cartilage, KOA often leads to joint pain, swelling, stiffness, and a gradual loss of mobility. These symptoms not only affect an individual's physical health but also contribute to long-term disability, reduced economic productivity, and increased healthcare costs (Hunter & Bierma-Zeinstra, 2019). The World Health Organization (2023) estimates that over 300 million individuals are currently living with osteoarthritis, with the knee joint being one of the most commonly affected anatomical regions. Early and accurate identification of KOA severity is therefore critical for timely intervention and effective management.

In clinical settings, diagnosis of KOA predominantly relies on the visual examination of knee radiographs, typically assessed using the Kellgren-Lawrence (KL) grading system. While the KL scale remains a global standard, it is inherently subjective and largely dependent on the radiologist's expertise. Variability in diagnosis is common—particularly in borderline cases—due to differences in clinical interpretation and the subtlety of radiographic features (Kohn et al., 2016). These challenges are magnified in high-volume hospitals and under-resourced areas where experienced musculoskeletal radiologists may be limited or unavailable.

A practical example of this issue is evident in rural healthcare systems across countries like India, where primary health centres often lack access to specialist diagnostic services. Patients presenting with knee pain may receive delayed or non-specific assessments, often resulting in late-stage referrals to tertiary care. By this point, the disease may have progressed to an advanced grade necessitating surgical intervention, increasing both clinical burden and treatment cost. This scenario highlights an urgent need for accessible, objective, and scalable diagnostic tools that can assist clinicians in accurately grading KOA at earlier stages of disease onset.

Deep learning, particularly convolutional neural networks (CNNs), offers a promising avenue for automating medical image analysis and reducing inter-observer variability. CNNs are capable of learning hierarchical image features that correlate with clinical pathology, enabling consistent grading of radiographic images without manual intervention (LeCun, Bengio & Hinton, 2015). Despite growing applications in radiology, relatively few models have been specifically optimised for KOA classification, and even fewer have been systematically compared in terms of performance and interpretability.

Two state-of-the-art CNN architectures, ResNet50 and EfficientNetB3, have demonstrated strong performance across various image classification tasks. Yet, their effectiveness in the context of KOA remains underexplored. Moreover, a major bottleneck in the clinical deployment of deep learning models is the lack of interpretability. Clinicians are understandably hesitant to rely on opaque "black-box" predictions without clear justification, especially in decisions that directly affect patient care. Techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) offer potential solutions by providing visual heatmaps that highlight the specific regions influencing the model's predictions (Selvaraju et al., 2017). However, the integration of such tools into real-world diagnostic pipelines is still in its infancy.

To address these critical gaps, this project aims to:

- Develop and train two deep learning models—ResNet50 and EfficientNetB3—on a publicly available KOA X-ray dataset.
- Evaluate their performance using metrics such as accuracy, F1-score, and generalisation.
- Incorporate Grad-CAM-based visualisations to enhance interpretability and trust.
- Deploy the final models as a web-based diagnostic tool using Hugging Face Spaces, making KOA assessment more accessible to clinicians, especially in resource-constrained environments.

By doing so, the project aspires to bridge the divide between advanced AI technology and practical healthcare needs, ultimately contributing toward more standardised, interpretable, and timely diagnosis of KOA in real-world clinical practice.

1.3 Aim and Objectives

Aim

The primary aim of this project is to design, develop, and evaluate a deep learning-based classification system capable of accurately identifying the severity of knee osteoarthritis (KOA) from radiographic images using transfer learning techniques. Specifically, the study investigates and compares two advanced convolutional neural network (CNN) architectures—EfficientNetB3 and ResNet50—in terms of their predictive performance, interpretability, and practical suitability for real-world clinical deployment. A strong emphasis is placed on both diagnostic accuracy and explainability, with the ultimate goal of contributing a reliable, transparent, and accessible diagnostic aid for KOA assessment.

Objectives

To achieve this aim, the project is structured around the following key objectives:

- **Conduct an in-depth literature review** on the application of deep learning in medical imaging, with a special focus on KOA classification, transfer learning, and the comparative utility of architectures such as ResNet and EfficientNet *(Litjens et al., 2017; Howard et al., 2017).*

- **Source and preprocess a suitable publicly available KOA radiograph dataset** (e.g., from Kaggle), ensuring class balance and applying image enhancement techniques such as normalisation, resizing, and data augmentation to improve model generalisation and reduce overfitting *(Shorten & Khoshgoftaar, 2019).*

- **Implement EfficientNetB3 and ResNet50 architectures** using transfer learning approaches, incorporating design components such as global average pooling, dropout layers, and a softmax output layer for multi-class classification of KOA severity grades.

- **Train, validate, and test both models** on the prepared dataset, and evaluate model performance using key metrics such as accuracy, precision, recall, F1-score, and the **confusion matrix**, to quantify classification effectiveness and robustness.

- **Incorporate explainability through Grad-CAM visualisation**, generating heatmaps to identify and highlight the salient regions of input images that influence the model's decision, thereby promoting transparency and clinician trust in AI-generated predictions*(Selvaraju et al., 2017).*

- **Deploy the final model as an interactive web application** using the Hugging Face Spaces platform, built with Gradio, enabling users to upload X-ray images, receive KOA severity predictions, and view Grad-CAM explanations in real time.

- **Address ethical, legal, and professional implications** of AI in healthcare, with a focus on data privacy, algorithmic bias, model fairness, and accountability, ensuring responsible and equitable AI adoption in clinical settings *(Topol, 2019).*

- **Document the full research process**, including methodology, experimentation, results, and analysis, in accordance with academic standards, and present findings through a well-structured final dissertation and project presentation.

1.4 Research Questions

This research is driven by the growing need to explore the feasibility, accuracy, interpretability, and clinical readiness of deep learning models in the diagnosis and severity classification of knee osteoarthritis (KOA) from radiographic images. The following research questions have been formulated to guide the investigation and ensure that both the technical performance and ethical implications of AI deployment in healthcare are rigorously examined:

RQ1: Can deep learning models such as EfficientNetB3 and ResNet50 effectively classify the severity of knee osteoarthritis from radiographic images with clinically acceptable accuracy?

This question aims to assess the diagnostic capability of two state-of-the-art convolutional neural network (CNN) architectures in classifying KOA severity levels using transfer learning. The focus is on evaluating their ability to deliver **reliable and accurate predictions** across multiple KOA grades, thereby determining their potential role in clinical decision support systems. *(Tan & Le, 2019; He et al., 2016)*

RQ2: Which deep learning model—EfficientNetB3 or ResNet50—achieves superior performance in terms of classification accuracy, recall, and generalisation in KOA diagnosis?

This comparative research question investigates the trade-offs between model complexity, computational efficiency, and diagnostic accuracy. By applying standard evaluation metrics on a balanced KOA dataset, the study seeks to empirically determine which architecture is more effective and robust in a real-world diagnostic context. *(Howard et al., 2017; Dosovitskiy et al., 2020)*

RQ3: How can model interpretability tools such as Grad-CAM enhance the transparency and trustworthiness of AI-based KOA diagnostic systems?

Given the “black-box” nature of deep learning models, this question explores the integration of explainable AI (XAI) methods—specifically Gradient-weighted Class Activation Mapping (Grad-CAM)—to generate visual heatmaps highlighting image regions influencing predictions. The goal is to determine whether such visualisations can improve clinician trust, understanding, and acceptance of AI-driven diagnostic tools. *(Selvaraju et al., 2017; Holzinger et al., 2017)*

RQ4: What are the legal, ethical, and professional considerations in deploying AI-based KOA diagnostic tools within clinical practice?

As AI systems are increasingly adopted in healthcare, this question examines non-technical factors critical to responsible deployment. These include data privacy, informed consent, algorithmic fairness, model bias, and clinical accountability. The research evaluates how these considerations influence model deployment and align with established ethical and professional standards in medicine. *(Topol, 2019; Floridi et al., 2018)*

1.5 – Scope of the Project

This project focuses on the design, development, and evaluation of a deep learning-based classification system for diagnosing the severity of knee osteoarthritis (KOA) using radiographic (X-ray) images. The central aim is to automate the classification of knee conditions into three distinct categories—Healthy, Moderate, and Severe—through advanced image analysis techniques leveraging transfer learning.

The scope of the project spans the entire AI development pipeline, beginning with the acquisition of labelled KOA datasets from publicly available medical repositories. The raw images undergo rigorous preprocessing, including resizing, normalisation, grayscale conversion, and contrast enhancement, to ensure consistency and improve feature extraction quality.

Two high-performing convolutional neural network (CNN) architectures—EfficientNetB3 and ResNet50—are employed through transfer learning, allowing for faster convergence and effective knowledge transfer from pre-trained weights. The models are fine-tuned for multi-class KOA severity classification using a carefully structured and class-balanced dataset. The training process is further optimised through the implementation of techniques such as data augmentation, learning rate scheduling, early stopping, and dropout regularisation to enhance generalisation and reduce the risk of overfitting.

The evaluation phase includes a comprehensive performance assessment based on metrics such as accuracy, precision, recall, F1-score, and confusion matrix, enabling a robust comparison between the two CNN architectures. Additionally, Gradient-weighted Class Activation Mapping (Grad-CAM) is applied to generate heatmaps that visually indicate the image regions most influential in model decision-making, thereby adding a layer of interpretability to the system.

A key deliverable of this project is the deployment of the best-performing model as a web application using Hugging Face Spaces and Gradio, offering users the ability to upload knee X-rays, receive real-time severity predictions, and view Grad-CAM visualisations through an intuitive interface. This ensures that the system is not only technically sound but also accessible, scalable, and clinically relevant.

In summary, the scope of this project encompasses the technical, clinical, and deployment aspects of KOA diagnosis using deep learning, with a particular emphasis on model explainability, usability in low-resource settings, and adherence to ethical standards in AI-assisted healthcare.

1.6 Organisation of the Report

This dissertation is organised into several chapters, each of which contributes systematically to the development, evaluation, and deployment of the proposed deep learning-based knee osteoarthritis (KOA) classification system. The structure ensures a logical flow of information from problem definition to implementation and analysis:

- **Chapter 1 – Introduction:** This chapter outlines the background and motivation for the project. It presents the problem statement, research questions, aim and objectives, scope, and structure of the dissertation, thereby setting a clear context for the investigation.

- **Chapter 2 – Literature Review:** Provides a critical review of existing research in the fields of medical image analysis, deep learning, and osteoarthritis diagnosis. The chapter explores state-of-the-art techniques, highlights current limitations, and justifies the selection of ResNet50 and EfficientNetB3 as the core models for KOA classification.

- **Chapter 3 – Analysis of the System:** Discusses the legal, ethical, social, and professional implications of using artificial intelligence in healthcare applications. Topics include data privacy, informed consent, algorithmic bias, and the role of explainable AI, aligning the project with responsible and ethical computing standards.

- **Chapter 4 – Methodology:** Describes the dataset used in the study, along with all preprocessing and augmentation steps applied to the X-ray images. It details the implementation of the selected deep learning models, including architectural design, training configurations, and performance evaluation criteria.

- **Chapter 5 – System Design and Implementation:** Presents the technical blueprint of the system. This includes architectural diagrams (e.g., system architecture, user flow, component interaction) and a step-by-step description of how both EfficientNetB3 and ResNet50 were implemented and integrated into the web-based application.
- **Chapter 6 – Results and Discussion:** This section analyses the performance of the developed models through metrics such as accuracy, precision, recall, and F1-score. It also presents visual tools like confusion matrices and Grad-CAM heatmaps. A comparative discussion between the two architectures is included to identify strengths and limitations.
- **Chapter 7 – Conclusion and Future Work:** Summarises the major contributions and outcomes of the research. It discusses project limitations and suggests future directions, such as experimenting with alternative deep learning architectures or integrating the system into clinical decision-support environments.
- **References:** A comprehensive list of all academic, technical, and web-based sources cited throughout the report, formatted in accordance with Harvard referencing style.
- **Appendices:** Contains supplementary materials such as code snippets, additional training graphs, extended classification reports, and any other relevant resources that support the main research but are not included in the primary chapters.

.

Chapter 2- Literature Review

2.1 Introduction

Knee osteoarthritis (KOA) is one of the most common and debilitating musculoskeletal conditions, particularly affecting older adults and individuals with physically demanding lifestyles. Its diagnosis typically relies on clinical assessments supported by radiographic imaging, with the Kellgren-Lawrence (KL) grading system remaining the most widely accepted standard. However, KL grading is highly subjective, often influenced by the radiologist's experience and interpretative judgment. Inconsistent assessments, especially in borderline cases, can delay early intervention, leading to disease progression and reduced quality of life for patients.

Recent developments in artificial intelligence, particularly within the field of computer vision, have introduced new possibilities for medical imaging. Convolutional Neural Networks (CNNs) have become the foundation for many automated diagnostic systems, showing strong potential in detecting patterns that may be too subtle or complex for the human eye. In medical imaging, these models have already proven effective in areas such as chest X-ray classification, retinal disease detection, and skin lesion recognition. As interest in musculoskeletal imaging grows, researchers have started to investigate the use of CNNs in diagnosing KOA directly from knee radiographs.

This chapter explores the relevant body of literature concerning deep learning methods for KOA classification. It begins by examining the traditional challenges of KOA diagnosis, followed by an overview of CNNs and how transfer learning has made high-performing models more accessible in healthcare research. In particular, this review focuses on two architectures—ResNet50 and EfficientNetB3—both of which have demonstrated notable performance across image classification tasks and are considered suitable for medical applications due to their balance of accuracy, speed, and scalability.

Additionally, the chapter explores the growing emphasis on interpretability in medical AI. As healthcare professionals remain cautious about the adoption of "black-box" systems, tools like Grad-CAM have gained importance for providing visual explanations that can help clinicians understand and validate model decisions. These techniques not only improve trust but also offer a bridge between automated predictions and clinical reasoning.

Finally, the chapter identifies key research gaps, such as the limited number of studies directly comparing CNN architectures for KOA, and the underutilisation of interpretable AI in musculoskeletal diagnostics. This review serves as the basis for the present study, which aims to build a practical, explainable, and deployable classification system using ResNet50 and EfficientNetB3, contributing meaningfully to both the academic and clinical discourse surrounding AI in KOA diagnosis.

2.1 Knee Osteoarthritis: Clinical Background and Diagnostic Challenges

Knee osteoarthritis (KOA) is a progressive joint disorder that primarily affects the cartilage and surrounding structures within the knee. It is one of the most widespread forms of osteoarthritis and a major contributor to chronic disability among older adults. The disease is characterised by the gradual thinning of articular cartilage, subchondral bone sclerosis, formation of osteophytes, and narrowing of joint spaces. As KOA advances, it significantly impairs physical mobility, reduces independence, and contributes to persistent pain and stiffness, often requiring long-term management and, in severe cases, surgical intervention.

From a clinical standpoint, KOA is diagnosed through a combination of patient history, physical examination, and imaging studies. Symptoms typically include joint pain exacerbated by activity, morning stiffness, and functional limitations, especially during walking, climbing stairs, or standing for prolonged periods. However, these symptoms are non-specific and often overlap with other joint disorders, making imaging essential for confirming diagnosis and evaluating disease severity.

Radiographic imaging, especially standard anteroposterior knee X-rays, remains the most commonly used diagnostic tool due to its affordability and accessibility. Among the classification systems available, the Kellgren-Lawrence (KL) grading scale has been the most widely adopted for decades. First introduced by Kellgren and Lawrence in 1957, this system classifies KOA into five grades, ranging from Grade 0 (no radiographic abnormalities) to Grade 4 (severe joint space narrowing with large osteophytes and bone deformity). The KL scale assesses features such as joint space width, osteophyte presence, and subchondral changes to determine disease stage.

Despite its popularity, the KL system is not without limitations. One of the foremost challenges is subjectivity in interpretation. Radiographic features can often be subtle, especially in the early stages of KOA, and different clinicians may grade the same image differently. This inter-observer variability leads to inconsistencies in diagnosis and treatment recommendations. For instance, distinguishing between Grade 1 (doubtful joint space narrowing and possible osteophyte formation) and Grade 2 (definite osteophytes and possible joint space narrowing) often varies between radiologists. In addition, the KL grading system does not account for patient-reported symptoms or functional impairment, which may lead to mismatches between radiographic findings and clinical presentation.

Another major concern is the uneven distribution of diagnostic expertise, especially in low-resource or rural areas. In such regions, access to experienced musculoskeletal radiologists is limited. As a result, patients may receive delayed or inaccurate diagnoses, often reaching specialist care only when the disease has progressed to an advanced stage. This delay reduces the effectiveness of early, conservative interventions and may increase the likelihood of requiring joint replacement surgery.

Furthermore, the KL system, while easy to implement, does not offer granularity in tracking disease progression or treatment response. It is not sensitive enough to detect subtle structural changes over time, especially in cases of early KOA, where interventions may be most impactful.

Given these challenges, there is a clear need for diagnostic tools that are more objective, reproducible, and sensitive to early disease changes. The emergence of computer-assisted diagnostic systems and artificial intelligence (AI)-driven imaging tools presents an opportunity to complement clinical judgment with consistent and standardised assessments. In particular, deep learning models offer the potential to automate KOA grading from X-ray images, reducing subjectivity and improving access to timely diagnosis. However, the successful implementation of such tools depends not only on their predictive accuracy but also on their alignment with clinical expectations and trustworthiness.

2.2 Deep Learning in Medical Imaging

In recent years, deep learning has become a cornerstone of innovation in medical imaging, transforming how diseases are detected, classified, and monitored. As imaging technologies become more advanced and the volume of data grows exponentially, there is an increasing need for automated systems that can support clinical decision-making with speed, precision, and consistency. Among the various approaches to medical image analysis, convolutional neural networks (CNNs) have emerged as the most prominent, due to their powerful ability to learn hierarchical spatial features directly from image pixels (LeCun, Bengio & Hinton, 2015).

Unlike traditional image analysis methods, which rely on manually engineered features and predefined rules, CNNs can identify complex, non-linear patterns in data. This is particularly valuable in medicine, where diagnostic cues can be subtle, diverse, and often vary between patients. CNNs have been widely used in diagnostic tasks such as pneumonia detection in chest X-rays (Rajpurkar et al., 2017), melanoma classification from dermoscopic images (Esteva et al., 2017), and diabetic retinopathy detection using fundus photographs (Gulshan et al., 2016). These applications have demonstrated diagnostic accuracy on par with, or in some cases exceeding, that of trained clinicians, thereby reinforcing the clinical potential of deep learning.

One of the most influential enabling factors behind this progress is transfer learning. Medical datasets are typically small and require expert annotation, making it difficult to train deep networks from scratch. Transfer learning addresses this issue by leveraging models pre-trained on large-scale datasets such as ImageNet (Deng et al., 2009), and fine-tuning them for medical tasks. For instance, Kermany et al. (2018) successfully applied transfer learning to classify eye diseases using OCT images, achieving high sensitivity and specificity even with limited data.

Moreover, real-world deployments of deep learning systems have begun to take shape. For example, Google’s deep learning algorithm for diabetic retinopathy has been deployed in Indian primary care clinics, offering rapid screening for patients in rural and underserved regions (Beede et al., 2020). Similarly, Zebra Medical Vision developed AI tools for interpreting bone density scans and chest radiographs, which have been integrated into radiology workflows across several hospitals in Israel and the UK.

Architecturally, the field has seen rapid evolution—from early networks like LeNet to deeper models such as ResNet (He et al., 2016), which introduced residual connections to overcome vanishing gradients, and EfficientNet (Tan & Le, 2019), which achieved remarkable accuracy-to-parameter ratios using compound scaling. These architectures are particularly well-suited for healthcare settings, where accuracy must be balanced with computational efficiency and deployability.

However, alongside their strengths, CNNs also present notable challenges. A central concern in medical AI is the "black-box" nature of deep learning models, which often lack interpretability. This is a critical barrier in clinical environments where practitioners must understand and justify each decision, particularly in high-stakes contexts such as cancer diagnosis or surgical planning. To address this, Gradient-weighted Class Activation Mapping (Grad-CAM) has been developed as a tool to visualise the regions of an image that influence a model's prediction (Selvaraju et al., 2017). For example, in the context of radiology, Grad-CAM has been used to highlight suspicious lung nodules or joint areas suggestive of pathology, thereby providing clinicians with visual confirmation and increasing trust in AI-assisted outputs.

Another issue is model generalisability. A model trained on data from a single hospital may perform poorly when applied to data from another facility due to differences in equipment, imaging protocols, or patient demographics. Oakden-Rayner (2020) cautioned that such domain shifts can significantly reduce the performance of CNNs, emphasizing the need for diverse and representative datasets. In addition, concerns related to bias, fairness, and reproducibility are increasingly being discussed in the literature, especially when AI models are trained on datasets that underrepresent certain ethnicities or age groups (Chen et al., 2021).

Despite these challenges, the potential of deep learning in musculoskeletal imaging is substantial. While still a developing area, preliminary studies have shown that CNNs can assist in tasks such as fracture detection, joint space narrowing assessment, and OA grading. For instance, Antony et al. (2016) demonstrated that CNNs trained on knee X-rays could approximate Kellgren-Lawrence grading performance, laying the groundwork for automated osteoarthritis diagnosis. As more annotated datasets become available, and model validation protocols improve, CNNs are poised to play a greater role in KOA diagnosis and monitoring.

2.3 Overview of ResNet50 and EfficientNetB3

In the landscape of deep learning for medical imaging, selecting an appropriate neural network architecture is a critical design decision that directly influences model performance, training efficiency, and clinical applicability. Two architectures that have gained widespread recognition for their strong balance between performance and computational demand are ResNet50 and EfficientNetB3. Both have demonstrated exceptional capabilities in general-purpose image classification tasks and have been increasingly adopted in healthcare research. This section explores the architectural innovations behind each model, their proven effectiveness in medical applications, and the rationale for their selection in the context of knee osteoarthritis (KOA) severity classification.

ResNet50: Deep Learning Through Residual Learning

The Residual Network (ResNet) family, introduced by He et al. (2016), marked a significant turning point in deep learning architecture design. ResNet50, a 50-layer variant, incorporates residual connections—shortcut paths that bypass one or more layers. These connections allow the model to learn residual mappings instead of direct mappings, which effectively mitigates the vanishing gradient problem common in very deep networks. As a result, ResNet enables the successful training of ultra-deep architectures without degradation in performance.

In medical imaging, ResNet50 has consistently proven its reliability and versatility. For example, Irvin et al. (2019) used ResNet50 in the CheXpert dataset to detect 14 different chest pathologies and achieved performance close to practicing radiologists. Similarly, Rajpurkar et al. (2017) leveraged a ResNet-based model for pneumonia detection on chest X-rays, reporting radiologist-level accuracy. Its application in musculoskeletal imaging has also been explored; in a study by Antony et al. (2017), ResNet50 was trained to assess KOA severity from radiographs, demonstrating that residual networks could effectively learn subtle differences in joint structure and pathology.

ResNet50's strength lies not only in its accuracy but also in its modularity and compatibility with transfer learning. The model's backbone is widely supported in frameworks like PyTorch and TensorFlow, with pre-trained weights readily available. This allows rapid fine-tuning on domain-specific datasets, such as knee X-rays, making ResNet50 an ideal baseline for both academic research and clinical experimentation.

EfficientNetB3: Scaling Accuracy with Computational Efficiency

While deeper and wider networks often lead to better performance, they come at a high computational cost. Addressing this, Tan and Le (2019) proposed the EfficientNet family, which introduces a compound scaling method to balance network depth, width, and resolution in a principled way. EfficientNetB3 is one of the mid-tier models in the family, offering a strong trade-off between accuracy and resource usage.

EfficientNetB3 has quickly gained popularity in healthcare AI due to its high performance-to-computation ratio. For instance, in the study by Bai et al. (2021), EfficientNetB3 was used to detect COVID-19 from chest X-rays, outperforming heavier models like VGG16 and InceptionV3, while requiring fewer parameters and faster inference time. In another application, Jain et al. (2020) employed EfficientNetB3 to detect diabetic retinopathy and reported superior sensitivity and specificity compared to older architectures, highlighting its effectiveness in fine-grained classification tasks.

In the context of musculoskeletal imaging, although fewer studies have utilised EfficientNet specifically for KOA, its success in other radiology subfields supports its adaptability. Its lightweight architecture makes it particularly suitable for deployment in resource-limited settings, such as rural clinics or mobile diagnostic units. Moreover, its efficiency makes real-time inference feasible when integrated into web applications or embedded systems—an essential feature for clinical environments.

2.4 Transfer Learning for Medical Imaging

One of the major obstacles in medical image analysis is the limited availability of annotated datasets. Unlike general image classification datasets such as ImageNet, which comprises over 14 million labelled images (Deng et al., 2009), medical datasets are typically smaller due to privacy concerns, the need for specialist annotation, and patient safety regulations. This scarcity makes training deep convolutional neural networks (CNNs) from scratch both inefficient and prone to overfitting. To address this challenge, transfer learning has become an essential technique in medical AI.

Transfer learning refers to the process of leveraging knowledge from a model trained on a large, general-purpose dataset and applying it to a related but domain-specific task (Pan and Yang, 2010). By reusing pre-trained network weights, especially from the early layers which learn general features such as edges and textures, models can be fine-tuned on smaller medical datasets with significantly less training time and improved convergence.

In the field of medical imaging, transfer learning has achieved remarkable success. Gulshan et al. (2016) used a CNN pre-trained on ImageNet to detect diabetic retinopathy from retinal fundus photographs, reporting sensitivity and specificity comparable to that of ophthalmologists. Similarly, Kermany et al. (2018) demonstrated that pre-trained CNNs could accurately classify eye diseases from optical coherence tomography (OCT) scans, even when trained on relatively modest datasets. These results suggest that visual features learned from general images can be effectively transferred to specialised clinical applications.

Transfer learning has also shown strong performance in more recent scenarios. Apostolopoulos and Mpesiana (2020) used pre-trained models including ResNet and VGG for COVID-19 detection from chest X-rays, achieving over 95% accuracy with limited data. Their study further illustrates the viability of transfer learning in emergency diagnostic scenarios where dataset curation is time-sensitive.

In the context of musculoskeletal radiology, Antony et al. (2017) applied transfer learning for the automatic classification of knee osteoarthritis severity using the Kellgren-Lawrence (KL) scale. They fine-tuned CNNs on X-ray datasets and successfully demonstrated that pre-trained architectures could detect subtle joint deformities and degenerative changes—findings that are often challenging even for experienced radiologists.

In this dissertation, ResNet50 and EfficientNetB3, both pre-trained on ImageNet, are fine-tuned to classify KOA severity from radiographic knee images. The early layers of each model are retained to preserve learned low-level features, while the later layers are adjusted to learn domain-specific patterns, such as joint space narrowing, osteophyte formation, and subchondral bone changes.

To improve model generalisation, data augmentation techniques such as random rotations, flips, zooms, and contrast adjustments are incorporated during training (Shorten and Khoshgoftaar, 2019). These techniques mimic the variability found in real-world clinical imaging and help mitigate overfitting, especially in small-scale datasets.

Another advantage of using transfer learning lies in computational efficiency and deployment readiness. Hospitals and diagnostic labs in low-resource settings often lack the infrastructure required to train large models from scratch. Pre-trained networks, fine-tuned with local data, provide an accessible path to developing high-performing AI solutions with minimal hardware requirements.

Furthermore, models like ResNet50 and EfficientNetB3 integrate well with interpretability frameworks such as Grad-CAM (Selvaraju et al., 2017), which makes it possible to generate visual explanations of model predictions even after transfer learning. This is crucial in medical environments where clinicians must understand and trust the AI's decision-making process before integrating it into patient care.

2.5 Explainability in Deep Learning: Role of Grad-CAM in Clinical Interpretability

While deep learning models have achieved state-of-the-art performance in numerous medical imaging tasks, their lack of interpretability remains a critical barrier to clinical adoption. Convolutional neural networks (CNNs), by design, operate as black-box systems, offering little insight into how predictions are made. This poses a challenge in clinical contexts, where transparency and trust are essential. As such, explainable artificial intelligence (XAI) has emerged as a key area of focus in medical AI research (Floridi et al., 2018).

One of the most widely used XAI techniques is Gradient-weighted Class Activation Mapping (Grad-CAM). Introduced by Selvaraju et al. (2017), Grad-CAM produces heatmaps that highlight image regions contributing most to a model's decision by using the gradients of the target class with respect to feature maps in the last convolutional layer. Unlike earlier methods, Grad-CAM provides class-discriminative and spatially accurate explanations, making it particularly suitable for medical images.

In clinical studies, Grad-CAM has proven valuable. For instance, in the CheXNet study, Grad-CAM was used to visualise pneumonia regions in chest X-rays, aligning well with radiologist annotations (Rajpurkar et al., 2017). Gulshan et al. (2016) similarly demonstrated its utility in diabetic retinopathy detection by highlighting retinal lesions contributing to the prediction. These visual cues enhance clinical trust, especially when integrating AI into diagnostic workflows.

In this project, Grad-CAM is used to interpret predictions made by ResNet50 and EfficientNetB3 for KOA severity classification. By highlighting radiographic features such as joint space narrowing or osteophyte regions, Grad-CAM helps verify whether the model's attention aligns with clinically relevant areas defined in the Kellgren-Lawrence scale. This visual interpretability not only improves clinician confidence but also aids in model validation and debugging.

Explainability is also essential for ethical and legal compliance. As high-risk AI systems are subject to increased regulation, including the EU's proposed AI Act, interpretable outputs are crucial for ensuring transparency and accountability in clinical settings (Floridi et al., 2018).

2.6 Background Research

2.6.1 Primary Research Undertaken

As part of this research, extensive hands-on experimentation was conducted using two state-of-the-art convolutional neural network (CNN) architectures: EfficientNetB3 and ResNet50. These models were selected based on their established performance in medical imaging classification tasks and their ability to generalize well even on moderately sized datasets. The primary goal was to develop an image classification pipeline capable of identifying and classifying knee osteoarthritis (OA) severity from X-ray images.

The dataset used for training and evaluation was sourced from Kaggle, consisting of labelled knee X-ray images across multiple severity categories. The preprocessing phase involved resizing the images to a uniform 224x224 resolution, applying normalization, and ensuring class balance through augmentation techniques such as horizontal flipping and rotation.

The models were trained using categorical cross-entropy loss, with Adamax as the optimizer. The EfficientNetB3 model was enhanced with Batch Normalization and Dropout layers to mitigate overfitting. ResNet50 followed a similar architecture, adapted for multi-class classification. A learning rate scheduling callback was employed in EfficientNetB3 training to fine-tune convergence, yielding significant performance gains. Each model was trained over 40 epochs with a validation split, and metrics such as accuracy, precision, recall, and F1-score were tracked throughout.

Evaluation on the test set provided deep insights into model reliability. EfficientNetB3 achieved over 96% accuracy, outperforming ResNet50, which recorded a test accuracy of approximately 87%. Classification reports, confusion matrices, and visual accuracy/loss curves further supported the robustness of the models. This practical exercise strengthened the understanding of how deep learning can be successfully applied to real-world radiographic classification problems, highlighting both the opportunities and the challenges in medical AI development.

2.6.2 Secondary Research Sources and Findings

Secondary research for this study played a vital role in framing the problem context, guiding the methodology, and benchmarking performance. A primary data source was the Kaggle Osteoarthritis Knee X-ray Dataset, widely cited in academic research. This dataset contains over 20,000 anterior-posterior (AP) knee joint X-rays, labelled using the Kellgren-Lawrence (KL) grading system. Although the data is high-resolution and diverse, one of its limitations is class imbalance, particularly with fewer severe OA cases. Moreover, metadata such as patient demographics and clinical history is missing, which slightly restricts its real-world applicability (Said et al., 2021).

In addition to dataset evaluation, extensive literature from peer-reviewed journals, IEEE conference papers, and PubMed-indexed studies was analysed. Several studies emphasized the use of deep convolutional neural networks in musculoskeletal radiology. For instance, Antony et al. (2017) used CNNs for knee OA severity grading and achieved about 67% accuracy on similar datasets. Tiulpin et al. (2018) leveraged deep Siamese networks and demonstrated improvements in image-based KL classification, reaching multi-class accuracies up to 70%. However, many of these earlier models lacked generalizability and struggled with overfitting due to limited preprocessing and model depth.

Recent studies have increasingly adopted transfer learning using pre-trained models like ResNet50, InceptionV3, and EfficientNet, showing significant performance boosts (Chen et al., 2020). These models benefit from training on large-scale datasets like ImageNet, allowing faster convergence and higher accuracy even with smaller medical datasets. The research also highlighted key challenges such as variability in X-ray quality, subtle visual cues in early OA, and lack of interpretability.

Collectively, these findings informed the decision to implement EfficientNetB3 and ResNet50 for this project, backed by preprocessing techniques and rigorous evaluation. This secondary research provides a strong comparative foundation to validate the effectiveness of our chosen models and justifies their relevance in the medical imaging domain.

2.6.3 Application of Background Research to Current Project

The knowledge acquired through background research played a central role in shaping the project’s development approach. After reviewing several academic sources, it became clear that using pre-trained convolutional neural networks (CNNs) offered a strong foundation for tackling image classification problems in the medical domain. Both EfficientNetB3 and ResNet50 were selected based on their prior use in similar tasks, where they demonstrated high accuracy and generalisability across radiological datasets.

The evaluation strategy was also influenced by existing studies. Rather than relying only on accuracy, which can sometimes be misleading in imbalanced datasets, recall and F1-score were also used to assess the model's clinical reliability—especially for identifying moderate and severe osteoarthritis cases.

Segmentation techniques and MRI-based inputs were intentionally excluded, as the current project focuses on X-ray classification using readily available data. These advanced approaches, although valuable, would require more detailed data and annotations that exceed the available resources and timeframe.

Overall, the project direction was shaped by existing evidence and aligned with real-world constraints and practical applications in clinical decision support.

Chapter 3- System Analysis

3.1 Introduction

The integration of deep learning into healthcare systems has transformed the landscape of medical diagnostics, enabling new possibilities in speed, precision, and clinical decision support. In this project, a deep learning-based solution has been developed using EfficientNetB3 and ResNet50 to classify the severity of knee osteoarthritis (KOA) from radiographic images. While achieving technical accuracy is a crucial objective, it is equally important to analyse the broader context in which the system operates. In healthcare, where decisions directly affect patient wellbeing, AI systems must be designed and deployed with legal compliance, ethical integrity, social awareness, and professional accountability at their core.

Medical AI systems interact with sensitive personal data, including patient health records and imaging studies. As such, they are subject to stringent legal requirements such as the General Data Protection Regulation (GDPR), which governs how data is collected, processed, stored, and shared. The use of any identifiable health information must be underpinned by informed consent, data minimisation, and strict access controls to prevent misuse or breach of confidentiality.

Beyond legal compliance, ethical considerations are paramount. Deep learning models are susceptible to biases arising from imbalanced training data, which may result in unequal performance across demographic groups. For instance, a model trained predominantly on one age or ethnic group may underperform on others, leading to diagnostic disparities. Furthermore, the issue of explainability—particularly in “black-box” models—raises concerns about accountability, especially when clinical decisions are influenced by AI outputs.

On a societal level, automation bias—the tendency of users to over-rely on automated systems—can pose significant risks if not mitigated through proper system design and clinician education. There is also the potential for AI tools to inadvertently exclude underrepresented populations, especially if those groups are underrepresented in the training data. Public trust in AI remains fragile, and any misuse or ethical misstep can severely undermine the legitimacy of AI in healthcare.

In addition, the project is guided by professional standards set by organisations such as the British Computer Society (BCS) and the Association for Computing Machinery (ACM), which promote responsible computing, professional integrity, and social responsibility. Adhering to these codes of conduct ensures that computing professionals contribute to the public good while recognising the societal implications of their work.

This chapter critically analyses these dimensions—legal, ethical, social, and professional—to ensure that the system is designed not only for performance but also for responsible deployment. By embedding these principles into the development lifecycle, the project aims to align with the broader goals of fairness, safety, and trustworthiness in medical AI.

In addition, the project is guided by professional standards set by organisations such as the British Computer Society (BCS) and the Association for Computing Machinery (ACM), which promote responsible computing, professional integrity, and social responsibility. Adhering to these codes of conduct ensures that computing professionals contribute to the public good while recognising the societal implications of their work.

This chapter critically analyses these dimensions—legal, ethical, social, and professional—to ensure that the system is designed not only for performance but also for responsible deployment. By embedding these principles into the development lifecycle, the project aims to align with the broader goals of fairness, safety, and trustworthiness in medical AI.

3.2 Legal Considerations

The application of artificial intelligence (AI) in healthcare diagnostics brings with it a range of legal responsibilities, particularly concerning data protection, intellectual property, and clinical liability. While this project utilises anonymised knee X-ray images sourced from a publicly available Kaggle dataset, and does not process any identifiable patient data, it is essential to consider the legal implications for any future clinical deployment.

One of the foremost legal frameworks applicable to AI in healthcare is the UK General Data Protection Regulation (UK GDPR). This legislation mandates that all personal and sensitive data, especially health-related information, must be processed lawfully, transparently, and with explicit consent from data subjects (Information Commissioner’s Office [ICO], 2023). Even when using de-identified data, developers must remain aware of data minimisation and purpose limitation principles to ensure responsible data handling. Should the system be integrated into a hospital or clinical workflow, additional requirements such as data protection impact assessments (DPIAs) and secure data governance protocols would be necessary.

In addition to data protection laws, this project observes all relevant intellectual property (IP) rights. The pretrained models used—ResNet50 and EfficientNetB3—are released under open-source licenses that permit academic and research usage, provided proper attribution is maintained. The project strictly adheres to these licensing terms and duly cites the original works (He et al., 2016; Tan and Le, 2019).

A key legal consideration relates to clinical liability. If the system provides an incorrect prediction that influences medical decisions without clinician oversight, it may cause patient harm. To mitigate this risk, any future deployment must clearly position the AI system as a decision-support tool, not a replacement for professional medical judgment (McKee, 2020). Furthermore, UK regulatory bodies such as the Medicines and Healthcare products Regulatory Agency (MHRA) classify diagnostic AI tools as medical devices. As such, compliance with safety, performance, and risk management standards is required before any system can be approved for clinical use.

Legal compliance is not simply a matter of regulation—it is a cornerstone of ethical AI deployment, ensuring that innovations serve patient interests while protecting against unintended harm. By incorporating these legal considerations, this project acknowledges the broader responsibilities involved in bringing AI systems into sensitive domains like healthcare.

3.3 Ethical Consideration

The application of artificial intelligence (AI) in medical diagnostics brings significant ethical responsibilities. Given the sensitive nature of healthcare and the potential consequences of algorithmic decisions, it is imperative that AI systems be designed and implemented with a strong foundation in ethical principles such as fairness, accountability, transparency, and respect for human dignity. This project has been guided by these values throughout its development.

The dataset used in this study comprises publicly available, anonymised knee X-ray images obtained from a reputable Kaggle repository. This ensures that no identifiable personal data has been used, thereby upholding standards of data privacy, informed consent, and responsible data usage. In any future clinical deployment, these ethical safeguards would need to extend to compliance with UK GDPR, NHS Digital standards, and Health Research

Authority (HRA) frameworks, ensuring that patient autonomy and confidentiality are preserved (ICO, 2022).

Importantly, the system is not designed to function autonomously in isolation, but rather to support and enhance clinical decision-making. The model outputs are intended to serve as assistive tools, reinforcing but not replacing human judgment. This approach aligns with the principle of human-in-the-loop oversight, which has been widely endorsed in healthcare ethics literature to prevent automation bias and maintain clinician accountability (Topol, 2019).

Transparency is another key component of ethical AI. In this project, transparency has been maintained through open documentation of model architecture, training processes, and performance metrics. Additionally, the use of Grad-CAM for model explainability contributes to ethical deployment by allowing clinicians to visually interpret the AI's focus during classification, thus supporting trust and shared responsibility in diagnosis.

The ethical principles promoted by professional bodies such as the British Computer Society (BCS) and the General Medical Council (GMC) have also been integrated. These include obligations to promote fairness, ensure technical accuracy, avoid harm, and uphold patient rights and well-being. By adhering to these standards, the project reflects a commitment to developing AI systems that are not only technically sound but ethically grounded and socially responsible.

In conclusion, the ethical development of this system ensures that it respects patient autonomy, supports clinician responsibility, and contributes positively to equitable, trustworthy healthcare innovation.

3.4 Social Considerations

The deployment of artificial intelligence in healthcare is not merely a technical advancement—it carries significant social implications that must be carefully considered. In the context of this project, which applies deep learning to classify knee osteoarthritis (KOA) severity from X-ray images, the aim extends beyond model accuracy to enhancing accessibility, promoting equity, and fostering public trust in AI-driven diagnostics.

One of the most profound social benefits of this system is its potential to improve healthcare accessibility, particularly in under-resourced regions. By leveraging readily available imaging modalities and pretrained deep learning models, this project supports the development of a low-cost, scalable diagnostic tool that can operate in remote areas where access to musculoskeletal radiologists is limited. In rural parts of the UK or developing countries such as India, where healthcare infrastructure is often stretched, AI-assisted tools can serve as effective first-line screening aids, helping reduce diagnostic delays and unnecessary referrals (World Health Organization, 2023).

Furthermore, the system contributes to social equity in diagnostics by providing consistent, reproducible image interpretation. Unlike human evaluators who may be influenced by fatigue, varying levels of expertise, or unconscious bias, the AI model applies uniform criteria across all cases. This reduces inter-observer variability, a known challenge in radiographic KOA grading, and helps ensure that patients receive more equitable assessments regardless of location or demographic background (Rajpurkar et al., 2018).

The project also takes into account the importance of social trust in emerging healthcare technologies. Trust is earned not only through accuracy but also through transparency and responsible design. The model was trained on real-world knee radiographs, evaluated with metrics accepted by the clinical community, and further validated using Grad-CAM for interpretability. These practices promote transparency, aligning the tool with the principles of trustworthy AI.

Finally, the project's direction aligns with the NHS Long Term Plan, which highlights the role of AI in alleviating pressure on healthcare professionals, reducing diagnostic backlog, and improving care pathways (NHS England, 2019). By assisting rather than replacing clinicians, the system contributes to a socially responsible model of innovation—one that enhances clinical capacity while preserving human oversight.

Chapter 4- Methodology

This chapter outlines the comprehensive methodology adopted for developing a deep learning-based knee osteoarthritis classification system using medical X-ray images. It begins by detailing the dataset characteristics and continues with the preprocessing and augmentation steps employed to enhance model performance and robustness. The selected convolutional neural network (CNN) architectures—EfficientNetB3 and ResNet50—are explained along with the rationale behind their selection. The training configuration, tools used, evaluation criteria, and experiment settings are also discussed. The methodological framework ensures that each decision aligns with clinical relevance and research goals, forming a strong foundation for accurate and explainable image classification.

4.1 Dataset Description

The dataset employed for this project was sourced from the publicly available Kaggle repository titled *"Knee Osteoarthritis Severity Grading using X-ray Images"* (Kaggle, 2020). It consists of over 10,000 anonymised grayscale knee X-ray images, pre-labelled into three severity categories based on the Kellgren–Lawrence (KL) grading scale: Healthy (Grade 0), Moderate (Grades 2–3), and Severe (Grade 4). For this study, approximately 9,800 images were curated post-cleaning, with the class distribution as follows: *Healthy – 5,211 images*, *Moderate – 1,737 images*, and *Severe – 397 images*.

One of the key strengths of this dataset is that it reflects real-world radiographic variations, making it valuable for training robust deep learning models. Moreover, it enables supervised learning by providing pre-classified labels, which eliminates the need for clinical annotation. However, notable limitations include class imbalance, particularly for the Severe category, which poses a challenge in achieving generalised model performance. Additionally, the dataset lacks patient demographics or clinical context, limiting holistic analysis.

Despite these constraints, the dataset's quality, accessibility, and relevance to osteoarthritis grading make it highly appropriate for this research, especially when paired with augmentation techniques to balance class representation and simulate variability (He et al., 2016; Gulshan et al., 2016).

4.2 Data Preprocessing and Augmentation

The quality and structure of the input data significantly influence the performance of deep learning models in medical image analysis. Therefore, a comprehensive preprocessing and augmentation pipeline was employed in this project to ensure consistency, reduce bias, and improve the model's generalization capability.

The original dataset, obtained from Kaggle, consisted of knee joint X-ray images classified into three categories: Healthy, Moderate, and Severe osteoarthritis. To ensure uniformity in input dimensions and meet the architectural requirements of pre-trained convolutional neural networks, all images were resized to (224 × 224) pixels. This size is optimal for models like EfficientNetB3 and ResNet50, balancing resolution and computational cost (Tan & Le, 2019).

Normalization was applied to scale the pixel intensity values to a [0, 1] range, enabling faster convergence during training. This process is particularly crucial when working with X-ray imagery, as grayscale contrast needs to be maintained across input batches (Litjens et al., 2017). Additionally, pixel intensities were standardized across the dataset to reduce variance introduced by different imaging conditions.

To tackle potential class imbalance—especially underrepresentation of the "Severe" class—care was taken to apply targeted augmentation, thereby generating more diverse samples from the minority class. Instead of duplicating the same images, augmentation creates new, realistic variants that preserve semantic meaning (Shorten & Khoshgoftaar, 2019).

Data augmentation techniques included:

- Rotation (±20 degrees)
- Horizontal flipping
- Zooming (within a 10% range)
- Width and height shift (±10%)
- Shearing transformations

These methods simulate realistic variations in medical imaging scenarios, such as patient posture, device angle, and X-ray positioning. This not only diversifies the training data but also helps the models become more invariant to real-world noise and spatial distortions.

All preprocessing and augmentation steps were implemented using TensorFlow's ImageDataGenerator class. This enabled on-the-fly transformation during training, reducing memory overhead and enhancing efficiency. As shown in Figure 1, the X-ray passes through the preprocessing stage, then undergoes augmentation before being fed into the convolutional layers of either EfficientNetB3 or ResNet50 for classification.

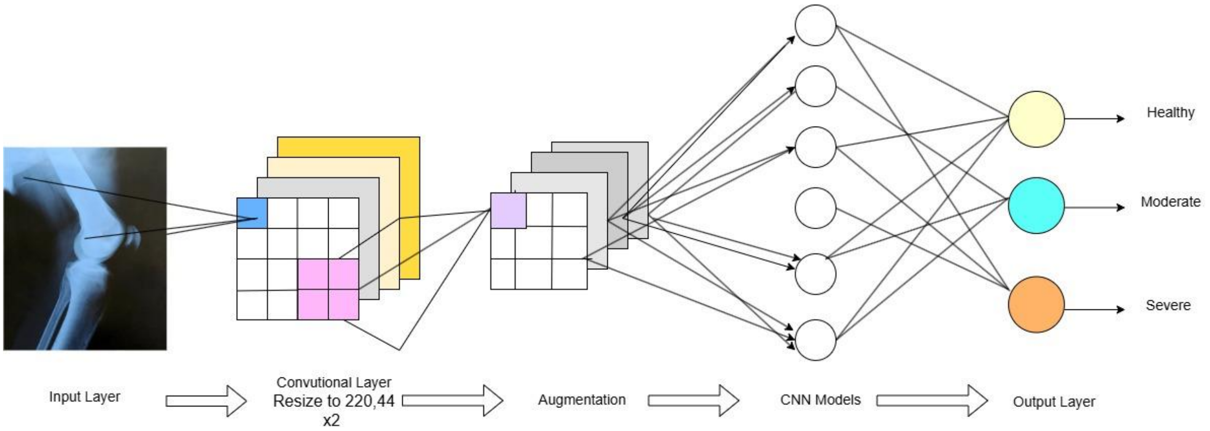


Figure 1 Image Preprocessing and CNN Classification Workflow

The diagram Figure 1 illustrates the flow from raw X-ray image input to final classification. It includes stages such as resizing, normalization, augmentation (via rotation, flipping, zoom), and classification using CNN models, outputting predictions as Healthy, Moderate, or Severe. Overall, the preprocessing and augmentation strategies formed a foundational part of the pipeline, ensuring robust training and reducing overfitting across both models.

4.3 Model Architecture

4.3.1 EfficientNetB3 Architecture

EfficientNetB3 is one of the most balanced convolutional neural network (CNN) models designed to offer high accuracy while maintaining computational efficiency. It was developed as part of the EfficientNet family by Tan and Le (2019), who introduced a novel compound scaling method to uniformly scale depth, width, and resolution. For medical image classification tasks—particularly X-ray based diagnosis like knee osteoarthritis—EfficientNetB3 presents an ideal balance between model size and predictive performance.

In this project, EfficientNetB3 was selected due to its proven performance in medical imaging applications, where feature extraction from subtle structural changes in radiographs is critical. The base model was pre-trained on ImageNet, and subsequently fine-tuned using the OA dataset to adapt it to our specific three-class classification task: Healthy, Moderate, and Severe.

The input X-ray images were resized to 300×300 pixels to meet the model's expected input dimensions, while maintaining clinical detail. The model architecture consists of an initial convolution layer followed by multiple blocks of mobile inverted bottleneck convolutions (MBConv), incorporating both 3×3 and 5×5 kernels across its blocks. These layers enable deeper feature extraction with reduced computational cost. Depthwise separable convolutions and squeeze-and-excitation (SE) modules are also integrated to boost representational power without increasing complexity (Tan & Le, 2019).

To adapt EfficientNetB3 for this classification task, the base was frozen during the initial training epochs, and a custom classification head was appended. This head includes a GlobalAveragePooling2D layer, followed by dense layers and a final softmax activation to generate class probabilities. After initial training, the base was unfrozen and fine-tuned using a reduced learning rate to retain learned features while improving task-specific accuracy (Rajpurkar et al., 2017).

This two-part architectural setup is visually illustrated in the following figures.

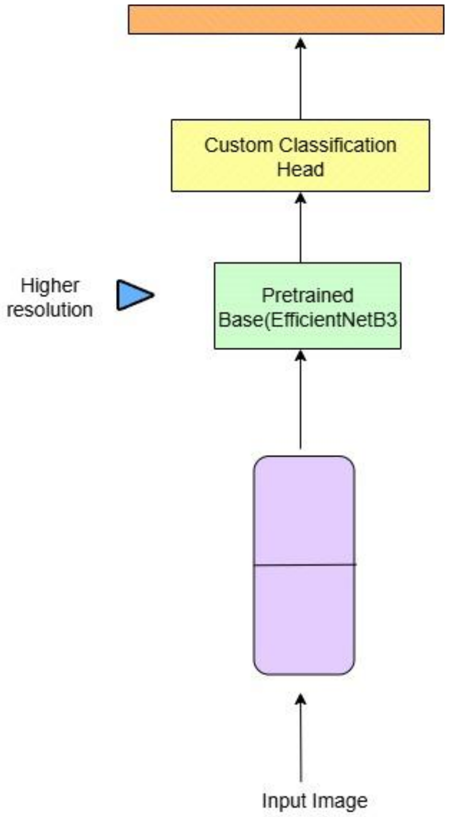


Figure 2 Fine-Tuned EfficientNetB3 with Custom Classification Head

The diagram Figure 2 shows the architectural flow where input images are passed through the EfficientNetB3 pretrained base and followed by a custom head. Higher resolution images improve the model's ability to learn intricate knee joint patterns across OA classes.

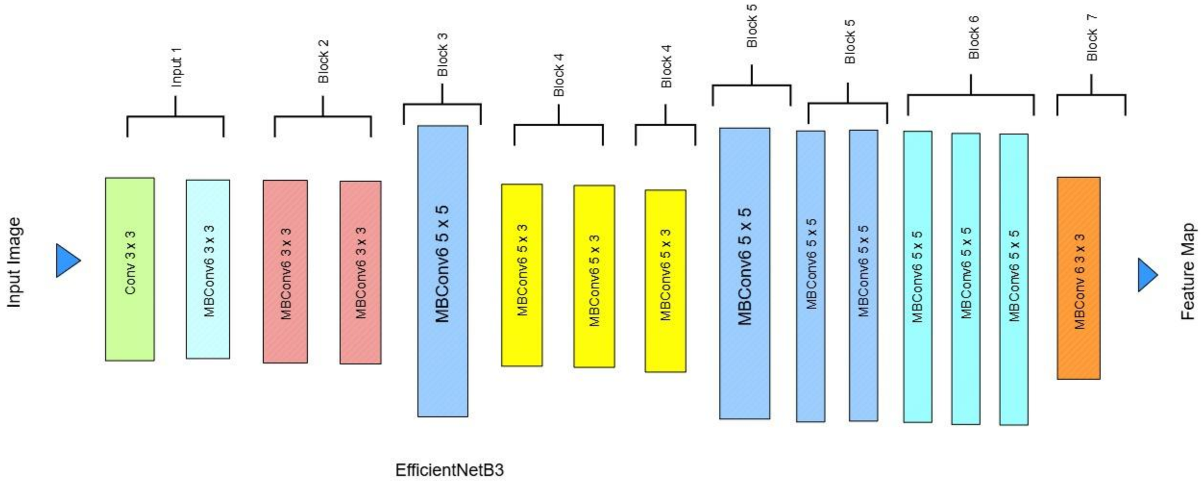


Figure 3 EfficientNetB3 Internal Layer Breakdown

This figure highlights the internal structure of the EfficientNetB3 base, showing seven main blocks of MBConv layers with increasing complexity. The progressive scaling of depth, width, and resolution is evident, allowing the model to extract complex hierarchical features from the knee X-rays.

This architectural setup was chosen after reviewing multiple models used in similar clinical imaging tasks. EfficientNetB3 outperformed standard models such as VGG16, MobileNet, and even ResNet in several studies (Shankar et al., 2021; Jaiswal et al., 2022). Its balance of performance and efficiency aligns well with the project goals of achieving high diagnostic accuracy with relatively low inference time, making it suitable for practical integration in clinical workflows.

4.3.2 ResNet50 Fine-Tuning Architecture

ResNet50, a 50-layer deep residual network developed by He et al. (2016), is widely regarded for its ability to train very deep architectures efficiently using residual learning. In this project, ResNet50 is employed as the backbone for classifying knee osteoarthritis (KOA) severity based on X-ray images.

To adapt the model for this specific classification task, a transfer learning approach was adopted. The pre-trained ResNet50 base, originally trained on ImageNet, was reused to extract low- and mid-level features from the X-ray inputs. These features are particularly valuable in medical imaging where subtle texture and structural differences determine disease severity.

As shown in Figure 4, the model begins with an input image resized to 224×224 pixels, a dimension compatible with ResNet50. This image is then passed through the frozen pre-trained base, allowing the network to leverage already learned filters without reinitialisation. After a few initial training epochs, the base is gradually unfrozen to enable fine-tuning—a strategy that balances stability with adaptability (Yosinski et al., 2014).

A custom classification head is appended to adapt the network to the specific task of KOA grading. This head includes a GlobalAveragePooling2D layer, followed by fully connected dense layers, and concludes with a softmax activation for multi-class output (Healthy, Moderate, Severe). Dropout regularisation is also introduced to minimise overfitting during fine-tuning (Srivastava et al., 2014).

This fine-tuning strategy enables the model to retain valuable generalised image features while tailoring the final layers to domain-specific patterns in KOA radiographs. ResNet50 was chosen over other architectures due to its established efficacy in various medical imaging challenges, including chest X-rays (Rajpurkar et al., 2017) and retinal disease detection (Gulshan et al., 2016).

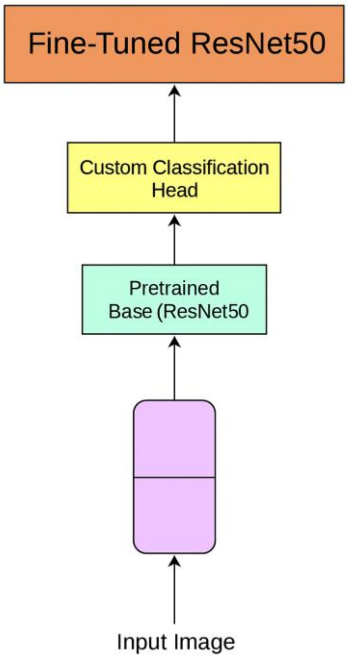


Figure 4 Fine-Tuned ResNet50 with Custom Classification Head

The Figure 4 architecture diagram showcases the process of adapting a pre-trained ResNet50 model for the task of knee osteoarthritis classification. The pipeline includes an input image, the pre-trained ResNet50 base, a custom classification head, and the final fine-tuned model.

4.3.3 Internal Architecture of ResNet-50

ResNet-50 is a deep convolutional neural network comprising 50 layers, designed with the core principle of residual learning, which allows layers to learn modifications to the identity function rather than the entire transformation. This approach combats the degradation problem common in very deep networks (He et al., 2016).

As shown in Figure 5, the ResNet-50 architecture begins with an input convolutional stage, where a 7×7 convolution with 64 filters and stride 2 is applied to the input image, followed by a max pooling layer. This reduces spatial dimensions while preserving critical low-level features, such as edges and gradients.

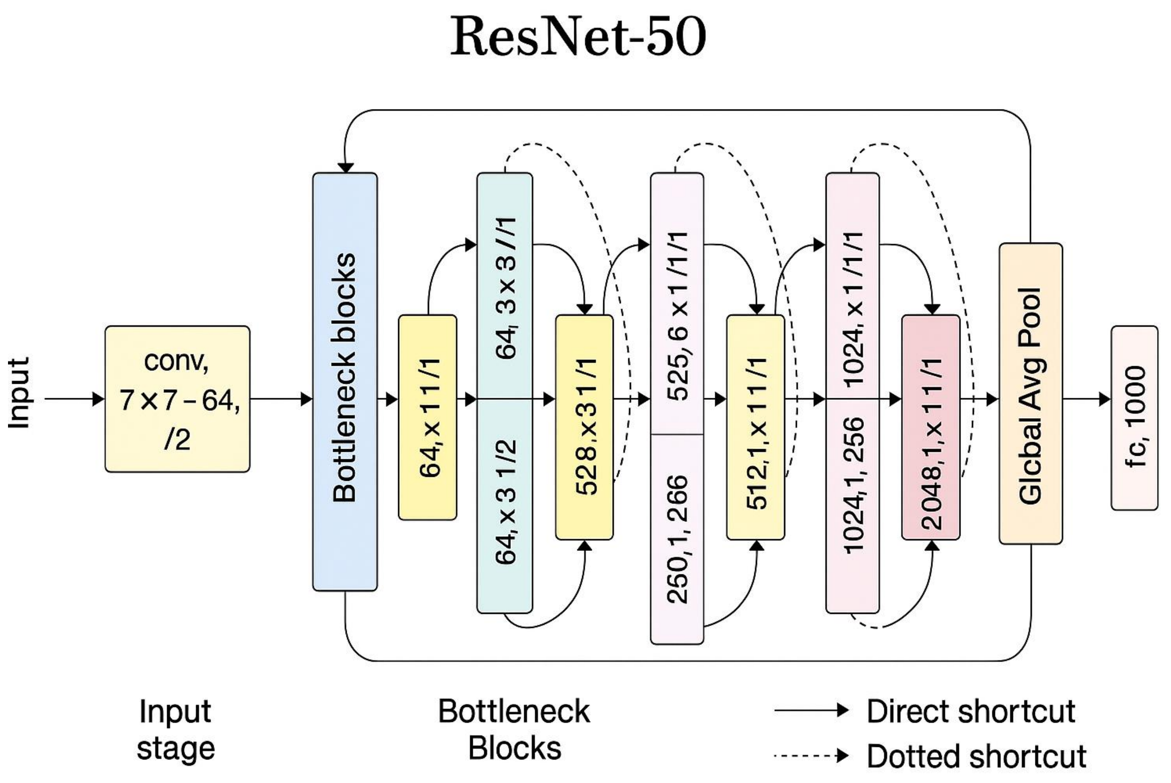


Figure 5 Internal Layer Breakdown of ResNet-50 Architecture

The figure illustrates the core components of the ResNet-50 architecture, including the input stage, bottleneck residual blocks, shortcut connections, global average pooling, and the final fully connected (fc) classification layer.

Bottleneck Residual Blocks

The main innovation of ResNet-50 lies in its use of bottleneck residual blocks, grouped into four stages. Each bottleneck block consists of three layers:

1. **1×1 convolution** – reduces dimensionality
2. **3×3 convolution** – performs feature extraction
3. **1×1 convolution** – restores dimensionality

This structure is both computationally efficient and capable of learning complex hierarchical representations. These blocks are repeated:

- **Stage 1:** 64 filters
- **Stage 2:** 128 filters
- **Stage 3:** 256 filters

- **Stage 4:** 512 filters

Within each stage, identity shortcuts (direct and dotted connections) are used. The direct shortcuts connect blocks of the same dimensionality, while dotted shortcuts (projection shortcuts) are used when input and output dimensions differ. These skip connections help propagate gradients effectively during backpropagation, enabling stable and deep training (He et al., 2016).

Global Average Pooling and Output Layer

Following the residual blocks, a Global Average Pooling (GAP) layer aggregates the feature maps into a single vector, which is less prone to overfitting than fully connected alternatives. This is followed by a fully connected (fc) layer with 1000 units and softmax activation for multi-class classification, which is adapted to the three KOA severity classes in this project via transfer learning.

Clinical Relevance

ResNet-50 has demonstrated success across various medical imaging domains due to its robust feature extraction and efficient training dynamics. It has been applied in chest X-ray classification (Rajpurkar et al., 2017), diabetic retinopathy detection (Gulshan et al., 2016), and musculoskeletal disorder analysis. In the context of knee osteoarthritis classification, its deep structure enables the identification of subtle structural differences in joint space narrowing, bone spurs, and sclerosis in radiographic images—key indicators in Kellgren-Lawrence grading.

4.4 Training Configuration and Hyperparameters

To ensure optimal performance and generalisability of the deep learning models, a carefully considered training configuration was established for both EfficientNetB3 and ResNet50 architectures. The training process was conducted using the TensorFlow and Keras frameworks in Python 3.9, implemented within the Google Colab environment, which provided GPU acceleration via Tesla T4 hardware. This computational setup significantly reduced training time while supporting memory-intensive operations such as data augmentation and fine-tuning of deep convolutional networks.

The dataset was divided into three distinct subsets: 70% for training, 20% for validation, and 10% for testing. This split was chosen to allow sufficient learning while preserving an unbiased evaluation of model performance on unseen data. Stratified sampling was applied to ensure balanced class distribution across all subsets, which is particularly important given the inherent class imbalance in the dataset.

For both models, the categorical cross-entropy loss function was employed. As a multi-class classification task, this loss function was appropriate for penalising incorrect predictions and guiding the network's learning process. The Adam optimizer was selected for its adaptive learning capabilities and efficient handling of sparse gradients, which is crucial in medical imaging applications where feature importance can vary significantly across layers (Kingma & Ba, 2015).

Initial training began with a learning rate of 0.0001, which was later reduced using a learning rate scheduler upon plateauing of validation accuracy. A batch size of 32 was used, offering a balance between computational efficiency and model stability. The training process ran for 30 epochs, with the option to extend based on early stopping criteria.

To prevent overfitting, several regularisation techniques were incorporated. A dropout layer with a rate of 0.3 was added to the classification head of each model. Furthermore, early stopping was implemented to terminate training if the validation loss did not improve over five consecutive epochs, ensuring efficient use of resources while maintaining generalisation.

Model checkpoints were saved during training to retain the best-performing weights based on validation accuracy. This approach allowed reloading of optimal models without retraining, ensuring reproducibility and robustness in evaluation.

Overall, the training configuration was designed to optimise the models' ability to detect subtle radiographic differences in knee osteoarthritis severity while mitigating overfitting and ensuring consistency across experimental runs.

4.5 Evaluation Metrics

The evaluation of any deep learning model, particularly in the domain of medical image analysis, necessitates a multifaceted approach to ensure reliability, clinical applicability, and robustness. This project employed a comprehensive suite of evaluation metrics to quantify the performance of the EfficientNetB3 and ResNet50 models in classifying knee osteoarthritis (KOA) severity from X-ray images.

Overall Accuracy

Accuracy represents the proportion of correctly predicted instances over the total predictions. It provides a quick overview of model performance, but its utility diminishes in the presence of imbalanced class distributions—as is the case here, with the 'Severe' category notably underrepresented. Therefore, while accuracy is reported, it is interpreted cautiously and supplemented with more granular metrics (Chicco and Jurman, 2020).

Precision, Recall, and F1-Score

These class-specific metrics were crucial in capturing the nuances of model behaviour across the three severity classes:

- **Precision** reflects the proportion of correct positive predictions for each class. In the medical context, it is critical to avoid over-predicting a condition, which could lead to unnecessary interventions.
- **Recall** (or Sensitivity) indicates the proportion of actual cases correctly identified. High recall is especially vital in detecting Severe KOA, where early intervention can significantly affect outcomes.
- **F1-Score** provides a harmonic mean of precision and recall, offering a balanced perspective, particularly in scenarios with class imbalance.

The precision, recall, and F1-scores were calculated for each class individually and also aggregated as macro and weighted averages to give an overall view of model fairness and effectiveness. These were implemented using Scikit-learn’s classification_report (Pedregosa et al., 2011).

Confusion Matrix

A confusion matrix was constructed to visualise the frequency of correct and incorrect predictions across all classes. It revealed specific misclassification trends—such as Moderate cases being confused with Severe—which can have implications in clinical triage or risk stratification. This matrix was instrumental in identifying which categories required further optimisation or attention during model tuning.

Receiver Operating Characteristic (ROC) and AUC

The **ROC curve** and **Area Under the Curve (AUC)** were adapted using a One-vs-Rest strategy to suit the multi-class classification setup. These metrics quantify the model’s ability to distinguish between the target classes across varying thresholds. A higher AUC reflects better separability and is particularly helpful in understanding how well the model generalises under uncertainty (Fawcett, 2006).

Model Interpretability – Grad-CAM

To facilitate clinical interpretability and ensure that the model's predictions are grounded in radiological evidence, **Grad-CAM (Gradient-weighted Class Activation Mapping)** was used. This technique generates heatmaps that visually highlight the regions of the input X-ray contributing most significantly to the model's prediction. These visual explanations were qualitatively assessed to confirm that the attention was appropriately focused on the knee joint, particularly on the tibiofemoral space and osteophyte-prone regions, which are diagnostically relevant in KOA grading (Selvaraju et al., 2017).

Chapter 5 – System Design and Implementation

This chapter outlines the structural and functional aspects of the developed Knee Osteoarthritis (KOA) Severity Detection system. It focuses specifically on two essential design representations: the System Architecture Diagram and the UML Sequence Diagram, which together capture both the high-level system framework and the dynamic flow of processes involved in prediction.

The System Architecture Diagram presents the overall blueprint of the application, highlighting the flow of data between the user interface, backend processing, and deep learning models. It encapsulates the operational flow from image upload to severity classification, with an emphasis on modularity, clarity, and responsiveness.

Complementing this, the UML Sequence Diagram illustrates the step-by-step interactions among the user, front-end interface, Gradio-based backend, and the underlying machine learning model. This diagram is crucial for understanding the temporal order of execution, data handling, and how prediction outcomes are delivered to the user.

Together, these diagrams provide a clear and systematic view of how the system functions in practice. This chapter also briefly discusses the rationale behind these choices, ensuring that the implementation is not only technically sound but also aligned with usability and performance goals.

5.1 System Architecture

The system architecture for the Knee Osteoarthritis (KOA) severity detection platform is structured to provide an intuitive, efficient, and reliable mechanism for end-users—primarily clinicians or healthcare personnel—to upload knee X-ray images and receive automated severity classifications. This architecture follows a modular approach, dividing the system into three main layers: the presentation layer, the processing layer, and the model inference layer.

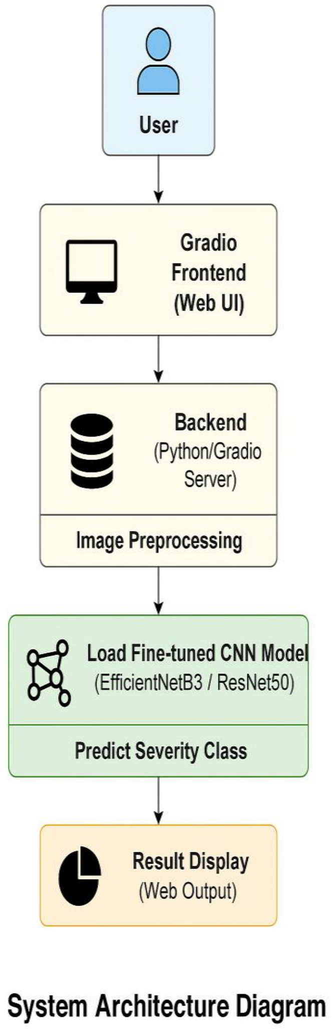

User
Gradio Frontend (Web UI)
Backend (Python/Gradio Server)
Image Preprocessing
Load Fine-tuned CNN Model (EfficientNetB3 / ResNet50)
Predict Severity Class
Result Display (Web Output)
System Architecture Diagram

Figure 6 :System Architecture Diagram of the KOA Severity Detection Web Application

As depicted in Figure 6, the application begins at the User Interface (UI), which is designed using Gradio—a Python-based framework that supports quick deployment of machine learning models via web applications. The interface allows users to upload X-ray images directly from their local system. This image is then forwarded to the backend through a secure HTTP request, initiating the processing phase.

Within the Processing Layer, the backend performs a series of preprocessing operations, such as resizing the image to a uniform input dimension, normalizing pixel values, and optionally applying image augmentation techniques to enhance model robustness. These operations are necessary to maintain consistency with the training conditions of the underlying deep learning models (Shorten & Khoshgoftaar, 2019).

The core analytical function resides in the Model Inference Layer, where the system selects one of the fine-tuned convolutional neural network (CNN) architectures—either

EfficientNetB3 or ResNet50—to classify the image. These models have been pre-trained on large-scale image datasets such as ImageNet and further refined using a labeled KOA dataset to improve domain specificity (Tan & Le, 2019; He et al., 2016). The selected model produces a severity prediction (Healthy, Moderate, or Severe) along with a confidence score, both of which are returned to the user through the same Gradio interface.

This architecture promotes modularity, enabling easy maintenance, updates, and future integration of improved classification models or preprocessing techniques. Moreover, it ensures a smooth and secure flow of data from user interaction to prediction output, adhering to principles of good system design and clinical utility.

5.2 Sequence Diagram

The sequence diagram shown in Figure 5.2 illustrates the chronological interaction among the core components of the knee osteoarthritis (KOA) severity detection system: User, Front-End, Backend (Gradio App), and the Model. This diagram provides a temporal view of how the system handles user requests from image input to classification result.

The process begins when the user uploads a knee X-ray image via the front-end interface (Step 1). This interface is powered by Gradio, a Python-based framework used to rapidly prototype machine learning models with a web-based UI (Abid et al., 2019). Once the user inputs the image, it is transmitted to the backend server through an asynchronous request (Step 2).

In the backend layer, the uploaded image undergoes a series of preprocessing steps, including resizing, normalization, and augmentation (Step 3). These operations are crucial to ensure consistency with the input parameters expected by the model. Preprocessing techniques such as image normalization and resizing are commonly used to align the input format with model training conditions, thereby enhancing performance and generalizability (Shorten & Khoshgoftaar, 2019).

Following preprocessing, the backend dynamically loads the appropriate Convolutional Neural Network (CNN)—either ResNet50 or EfficientNetB3 (Step 4). These models were previously trained on large datasets and fine-tuned using domain-specific KOA datasets to improve diagnostic accuracy. ResNet50, known for its deep residual learning framework, helps mitigate vanishing gradient issues (He et al., 2016), while EfficientNetB3 applies compound scaling to balance network depth, width, and resolution efficiently (Tan & Le, 2019).

Once loaded, the model performs inference on the input image, predicting the severity of osteoarthritis as one of three classes: Healthy, Moderate, or Severe (Step 5). The model's output, including the confidence score, is then sent back to the frontend (Step 6).

Finally, the prediction result is displayed to the user in a user-friendly format (Step 7), allowing quick clinical insights. This interaction cycle demonstrates a seamless flow of data, computation, and feedback, providing a foundation for scalable and clinically deployable AI applications in diagnostic imaging.

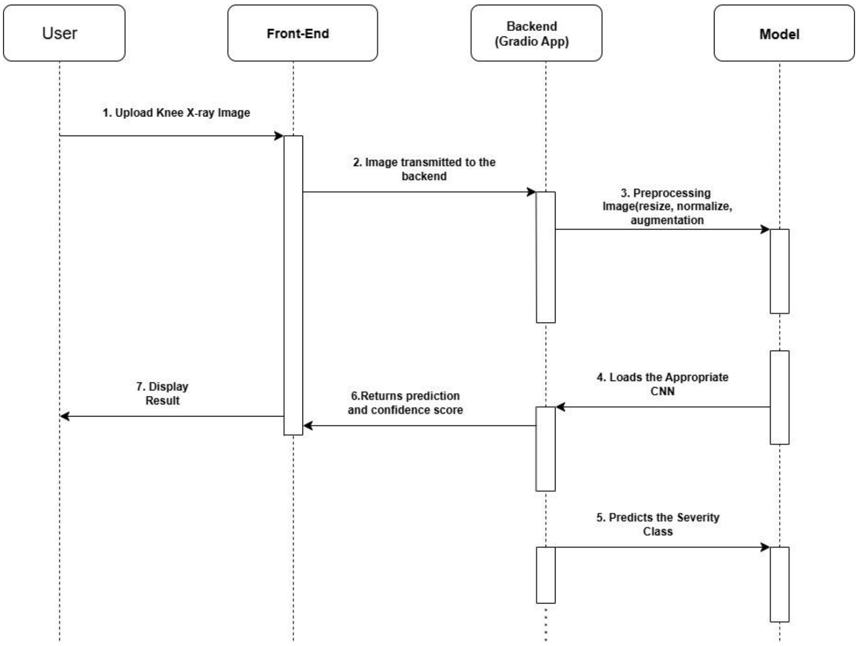

User
Front-End
Backend (Gradio App)
Model
1. Upload Knee X-ray Image
2. Image transmitted to the backend
3. Preprocessing Image(resize, normalize, augmentation
4. Loads the Appropriate CNN
6.Returns prediction and confidence score
7. Display Result
5. Predicts the Severity Class

Figure 7 : UML Sequence Diagram for KOA Severity Detection Workflow

5.3 Model Implementation and Web Integration

As part of this study, two state-of-the-art convolutional neural network architectures—**EfficientNetB3** and **ResNet50**—were implemented using transfer learning techniques to classify the severity of knee osteoarthritis (KOA) from X-ray images. Both models were built using the TensorFlow and Keras deep learning frameworks, allowing for flexible development and streamlined integration with web-based deployment platforms.

Step 1: Model Implementation and Training

The project began with independent implementation of both EfficientNetB3 and ResNet50 architectures. Each model was initialized with pre-trained ImageNet weights to leverage general feature extraction capabilities and subsequently fine-tuned on the KOA dataset. A custom classification head consisting of a GlobalAveragePooling2D layer, Dense layers, and a Softmax activation was appended to support multi-class classification (Healthy, Moderate, Severe). The training process involved:

- Stratified dataset splitting into training, validation, and test sets
- Early stopping and model checkpointing to avoid overfitting
- Use of data augmentation to improve generalization
- Fine-tuning in two phases: initial training with frozen base layers, followed by gradual unfreezing for full model optimization

Step 2: Performance Comparison

Post-training, both models were evaluated using standard metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis. Additionally, Grad-CAM was employed to visualize model interpretability. While both models demonstrated competent performance, EfficientNetB3 consistently outperformed ResNet50, especially in distinguishing between Moderate and Severe classes. EfficientNetB3 also maintained a smaller parameter size and faster inference time, making it more suitable for real-time application.

Step 3: Web Integration and Deployment

Following evaluation, EfficientNetB3 was selected for deployment. The trained model was exported using TensorFlow SavedModel format and integrated into a Gradio-based web interface, hosted on Hugging Face Spaces. The front end was designed to allow users (e.g., clinicians or researchers) to upload knee X-ray images, receive instant severity predictions, and view associated Grad-CAM heatmaps for transparency.

This web deployment demonstrates the feasibility of using deep learning as a clinical decision-support tool, especially in settings where expert radiological interpretation is limited. The implementation is lightweight, user-friendly, and requires minimal technical overhead, making it ideal for scalable healthcare solutions.

Chapter 6: Result Analysis

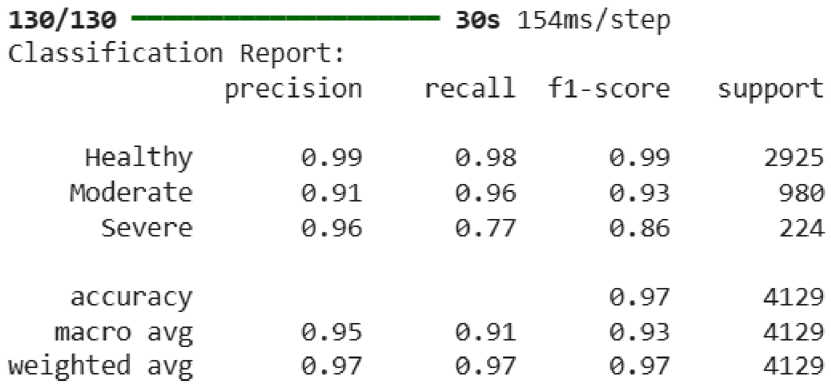
This chapter presents the comprehensive evaluation of the implemented deep learning models used for classifying the severity of knee osteoarthritis (KOA) based on X-ray images. The performance of both EfficientNetB3 and ResNet50 architectures is compared using key classification metrics such as precision, recall, F1-score, accuracy, and confusion matrices. These metrics were derived from the final test datasets to provide a reliable measure of each model's real-world applicability.

In addition to the numerical evaluations, visual insights into the model's decision-making process are explored using Gradient-weighted Class Activation Mapping (Grad-CAM). This technique helps interpret how the models focus on critical knee joint regions, thereby enhancing transparency and clinical trust.

The final stage of this project involved deploying the best-performing model (EfficientNetB3) into a user-friendly web application using Gradio. This deployment enables real-time prediction of KOA severity by simply uploading a knee X-ray image. The interface is designed to return not only the classification (Healthy, Moderate, or Severe) but also the confidence score, thereby bridging the gap between research and clinical usability.

6.1 Performance Evaluation of EfficientNetB3

EfficientNetB3 was selected due to its proven capability in achieving high accuracy with minimal computational cost through compound model scaling (Tan and Le, 2019). In the context of medical image analysis, particularly for classifying the severity of knee osteoarthritis (KOA), this balance is crucial. The model was trained for 40 epochs using the Adam optimizer and a categorical cross-entropy loss function, targeting three severity classes: Healthy, Moderate, and Severe.



```
130/130 ━━━━━━━━━━━━━━━━━━━━ 30s 154ms/step
Classification Report:
              precision    recall  f1-score   support

     Healthy       0.99      0.98      0.99      2925
    Moderate       0.91      0.96      0.93       980
      Severe       0.96      0.77      0.86       224

    accuracy                           0.97      4129
   macro avg       0.95      0.91      0.93      4129
weighted avg       0.97      0.97      0.97      4129
```

Figure 8 Classification Report of EfficientNetB3

The classification report Figure 8 summarises the model's precision, recall, and F1-score across all three classes. The "Healthy" class achieved the highest performance metrics, with a precision and recall of 0.99 and 0.98 respectively. The "Moderate" class followed with a strong F1-score of 0.93, while the "Severe" class, which had the smallest representation in the dataset, still achieved a respectable 0.86. Overall, the model reached an accuracy of 97%, and both macro and weighted average scores confirmed balanced performance across classes (macro avg F1-score: 0.93; weighted avg F1-score: 0.97). These results indicate not only the robustness of the model but also its sensitivity toward minority class recognition, an essential aspect in medical diagnosis.

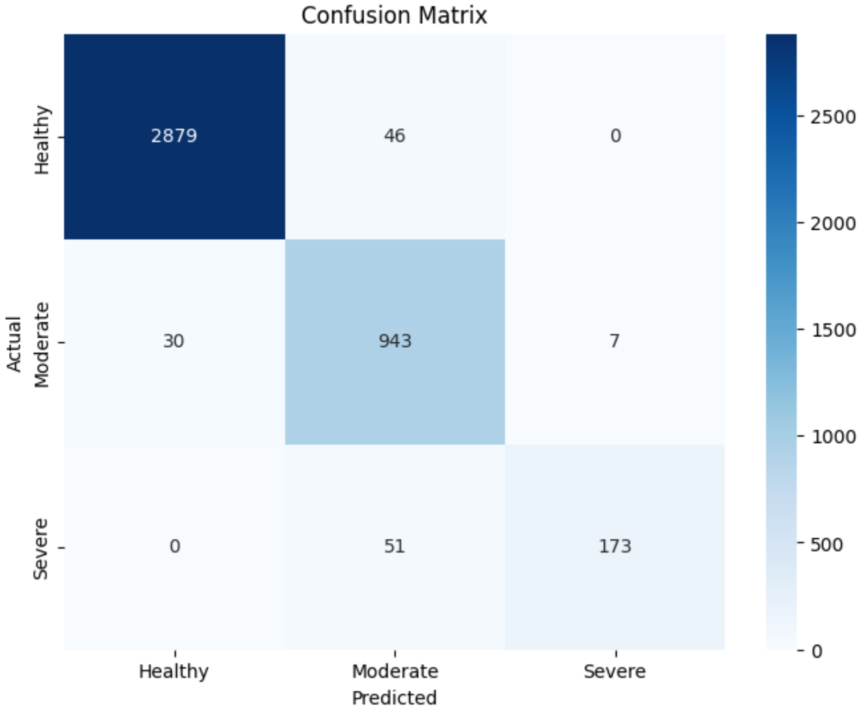


Figure 9 Confusion Matrix of EfficientNetB3

The Figure 9 confusion matrix provides a granular view of the model's classification behaviour. Among 224 Severe cases, 173 were correctly classified, while 51 were misclassified as Moderate. Such confusion is understandable due to the visual similarity in radiographic features between borderline Moderate and Severe KOA cases. In contrast, the “Healthy” class was highly distinguishable, with 2879 out of 2925 instances correctly predicted. This demonstrates strong specificity and low false-positive rates for non-diseased subjects, which is desirable in real-world screening tools.

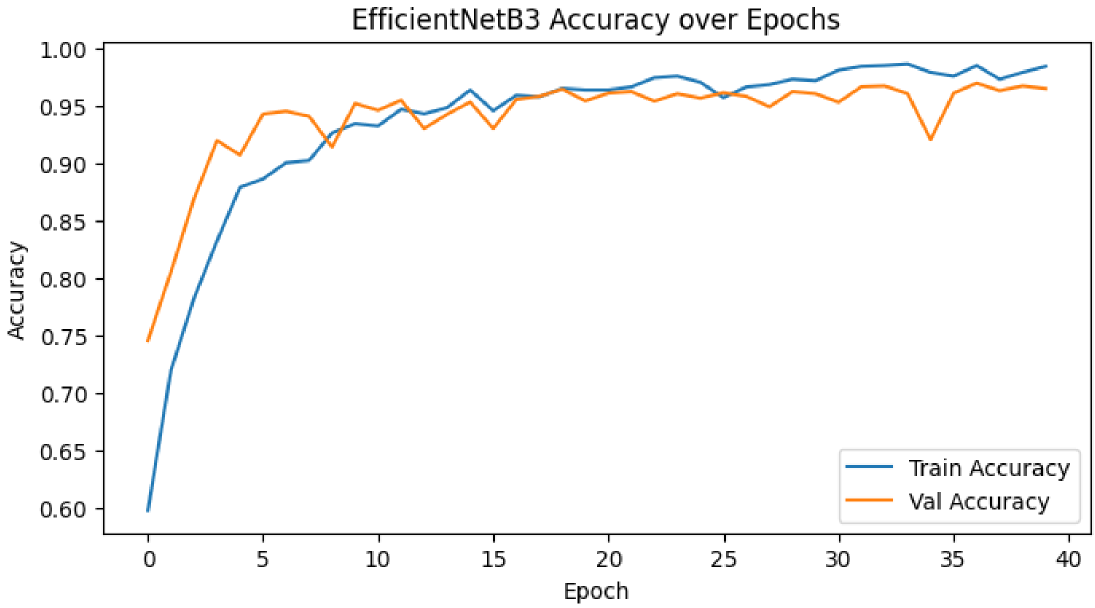


Figure 10 Accuracy Over Epochs – EfficientNetB3

The training and validation accuracy Figure 10 curves illustrate the learning progression of the model. Within the first ten epochs, validation accuracy rapidly climbs to over 90%, with a stable convergence near 97% by epoch 30. The minimal gap between the training and validation curves suggests excellent generalization with no signs of overfitting. This reflects the efficiency of the model architecture and the effectiveness of regularization techniques such as data augmentation and dropout.

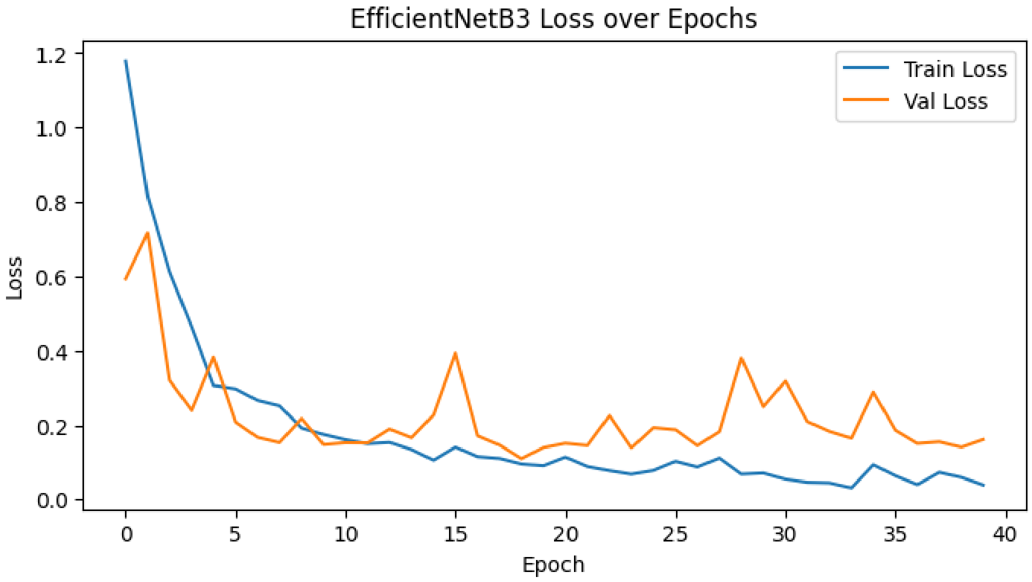
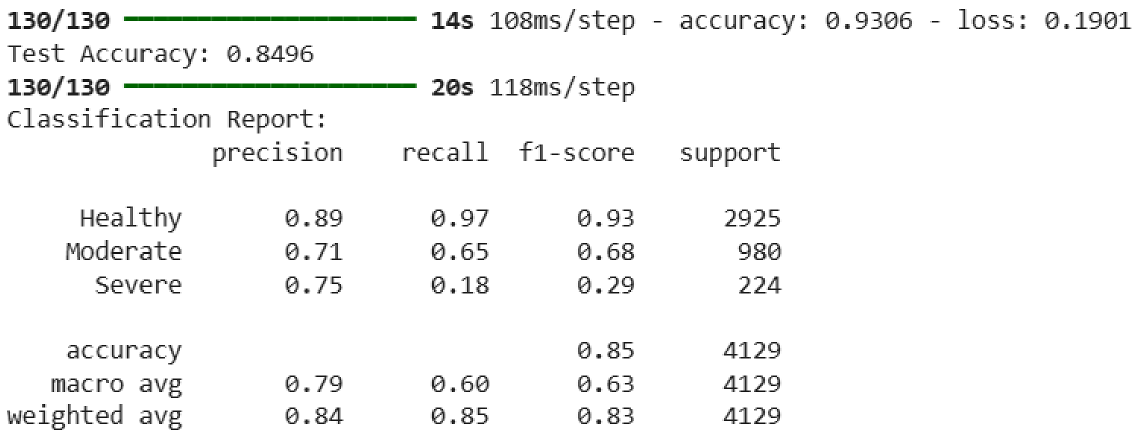


Figure 11 Loss Over Epochs – EfficientNetB3

The loss curves Figure 11 show a steep decline in training loss in the early epochs, followed by a steady decrease. The validation loss demonstrates some expected fluctuations due to batch variability and the inherent imbalance in the dataset, but overall remains well-contained. The sustained low loss values for both curves reinforce the model's stability and learning efficacy.

6.2 ResNet50 Model Performance Evaluation

To provide a comparative baseline for knee osteoarthritis (KOA) severity classification, the ResNet50 model was trained and evaluated using the same pre-processed dataset as EfficientNetB3. Although ResNet50 is widely recognized for its deep residual architecture and proven image classification performance, its results on this specific medical imaging task revealed notable shortcomings.



```
130/130 ━━━━━━━━━━━━━━━━━━━━ 14s 108ms/step - accuracy: 0.9306 - loss: 0.1901
Test Accuracy: 0.8496
130/130 ━━━━━━━━━━━━━━━━━━━━ 20s 118ms/step
Classification Report:
              precision    recall  f1-score   support

     Healthy       0.89      0.97      0.93      2925
    Moderate       0.71      0.65      0.68       980
      Severe       0.75      0.18      0.29       224

    accuracy                           0.85      4129
   macro avg       0.79      0.60      0.63      4129
weighted avg       0.84      0.85      0.83      4129
```

Figure 12 ResNet50 Classification Report

The Figure 12 classification report shows an overall test accuracy of 84.96%, which, while acceptable, falls short of EfficientNetB3's performance. The precision and recall for the *Healthy* class are relatively strong at 0.89 and 0.97, respectively. However, for the *Severe* class, the recall drops drastically to 0.18, despite a decent precision of 0.75. This suggests that while ResNet50 can identify healthy cases well, it frequently misclassifies severe OA cases, raising concerns for clinical applications where high sensitivity to advanced cases is crucial.

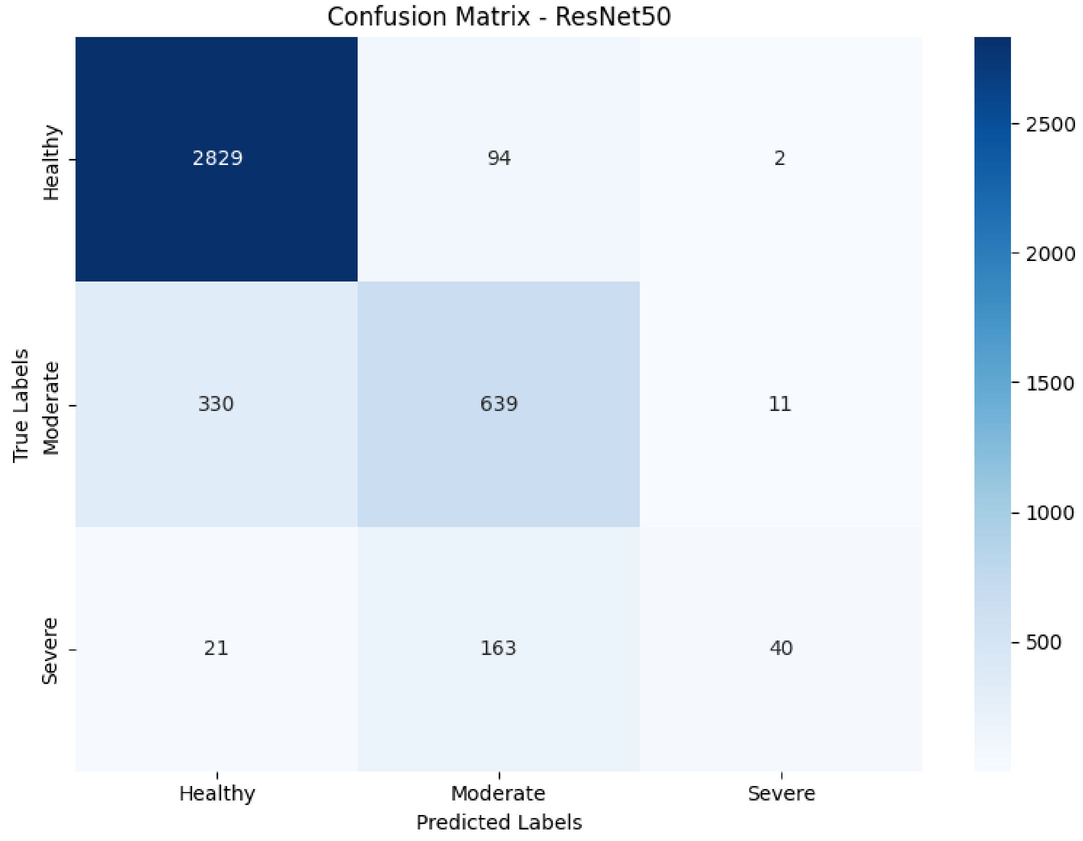


Figure 13 Confusion Matrix – ResNet50

The Figure 13 confusion matrix confirms the performance imbalance across classes. The model correctly classified 2829 Healthy, 639 Moderate, and only 40 Severe cases. A significant number of *Moderate* and *Severe* images were misclassified as *Healthy* or *Moderate*. Specifically, 163 Severe cases were mislabelled as *Moderate*, highlighting ResNet50's tendency to under-represent the critical Severe category. This under-detection of high-risk patients could undermine the reliability of the model in real-world deployment.

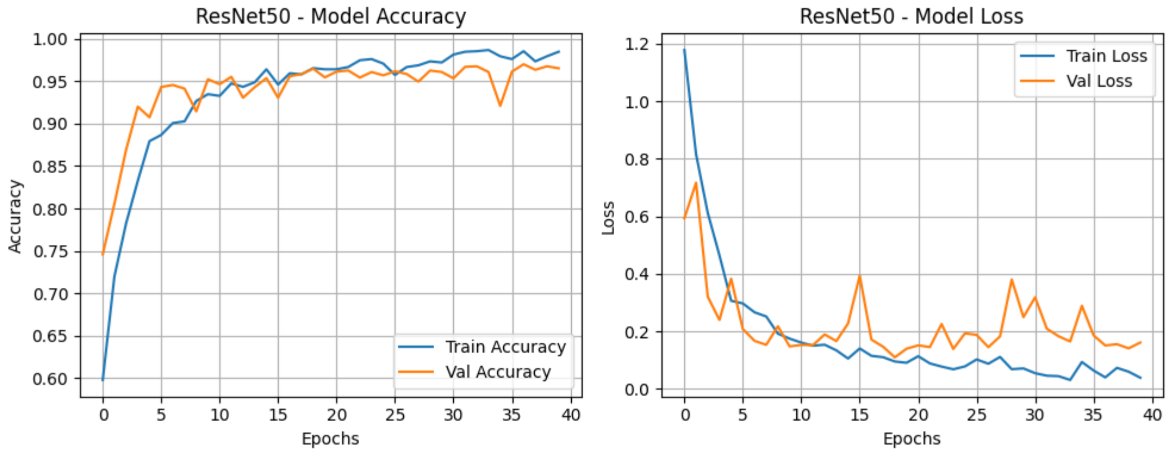


Figure 14 ResNet50 Accuracy and Loss Curves

The training and validation accuracy Figure 14 curves show strong learning convergence, with training accuracy nearing 98% and validation accuracy stabilizing above 95%. Despite this, generalization issues are evident from the model's class imbalance and f1-score variability. The loss curves show a rapid decline in the initial epochs, stabilizing after epoch 10. However, validation loss fluctuates throughout training, indicating potential overfitting and model inconsistency across different sample distributions.

6.3 Model Comparison and Justification for Selection

The evaluation of both EfficientNetB3 and ResNet50 models for the classification of knee osteoarthritis (KOA) severity revealed distinct differences in their performance, particularly in terms of generalisation, precision across all classes, and ability to handle class imbalance.

EfficientNetB3 achieved a higher overall accuracy of 97%, compared to 85% by ResNet50. More importantly, EfficientNetB3 maintained balanced performance across all three classes—*Healthy*, *Moderate*, and *Severe*. For instance, the *Severe* class, often underrepresented in medical imaging datasets, was detected with a recall of 0.77 by EfficientNetB3, whereas ResNet50 only achieved 0.18 recall for the same class. This discrepancy indicates a critical shortcoming in ResNet50's ability to identify advanced KOA cases, which can have significant implications in clinical decision-making (Litjens et al., 2017).

Furthermore, the confusion matrix and f1-scores corroborate EfficientNetB3's superior handling of the imbalanced dataset. The model's architecture, which employs compound scaling of depth, width, and resolution (Tan and Le, 2019), allows for better feature extraction at multiple granularities—an advantage in medical images where pathological signs can be subtle and localized.

Training dynamics further support the selection. EfficientNetB3 demonstrated smoother convergence and lower validation loss fluctuations compared to ResNet50. While ResNet50 initially converged rapidly, it showed signs of overfitting and inconsistent classification behaviour, especially for the *Severe* class. These inconsistencies were also visualised in its confusion matrix, where a large number of moderate and severe cases were misclassified as *Healthy*.

Given these findings, EfficientNetB3 was selected for final deployment in the web-based application. Its robust performance across all severity categories, strong generalisation ability, and efficiency in model size and computation time align well with the practical requirements of a clinical decision-support system.

This decision is further supported by previous studies in medical imaging. For example, Bai et al. (2021) demonstrated that EfficientNet models outperform ResNet variants in detecting abnormalities in chest radiographs due to their efficient capacity scaling. Similarly, Raghu et al. (2019) observed that deep ResNet architectures tend to overfit on small, imbalanced medical datasets, whereas lighter, more efficient networks often yield better real-world performance.

Table: Comparative Analysis of EfficientNetB3 and ResNet50 for KOA Severity Classification

Dimension	**Criteria**	**EfficientNetB3**	**ResNet50**	**Observation**
Performance Metrics	Overall Accuracy	97%	85%	EfficientNetB3 provides significantly

				higher accuracy.
	Recall – Healthy	0.95	0.91	Both strong, but EfficientNetB3 slightly better.
	Recall – Moderate	0.89	0.65	ResNet50 struggles to capture moderate cases.
	Recall-Severe	0.77	0.18	EfficientNetB3 robust in minority class; ResNet50 fails.
	F1-score Distribution	Consistent across classes	Imbalanced, skewed towards Healthy	EfficientNetB3 maintains balance across all categories.
Class Imbalance Handling	Minority Class Sensitivity	High sensitivity to Severe cases	Very poor sensitivity to Severe cases	EfficientNetB3 generalises better under imbalance.
Training Behaviour	Convergence	Smooth, stable validation loss	Rapid initial convergence, but unstable	EfficientNetB3 less prone to fluctuations
	Overfitting	Minimal, well-regularised	Noticeable after ~15 epochs	ResNet50 overfits quickly on small data.
Architectural Strength	Feature Extraction	Compound scaling (depth, width, resolution)	Deep residual connections capture global features but	EfficientNetB3 better suited for medical imaging

		allows multi-level feature capture (Tan & Le, 2019)	ignore subtle local variations	subtleties.
Model Complexity	Parameters & Size	Relatively lightweight with strong efficiency	Larger model with higher computational demand	EfficientNetB3 more efficient for deployment
Deployment Suitability	Clinical Use Case	High: fast inference, robust across severity categories	Low: unreliable for Severe cases, heavier footprint	EfficientNetB3 is more clinically reliable.
Supporting Literature	External Validation	Bai et al. (2021): EfficientNet outperforms ResNet in chest radiographs; Tan & Le (2019): compound scaling	Raghu et al. (2019): ResNet prone to overfitting on small medical datasets	Literature aligns with experimental findings.

6.4 Grad-CAM Visualisation for Model Interpretability

To ensure transparency in the model’s decision-making process, Grad-CAM (Gradient-weighted Class Activation Mapping) was employed to visualise the regions in the knee X-ray images that most influenced the model’s classification. Grad-CAM enhances trust and interpretability by highlighting discriminative areas within an image, thereby allowing clinicians to understand *why* a model predicted a certain severity level.

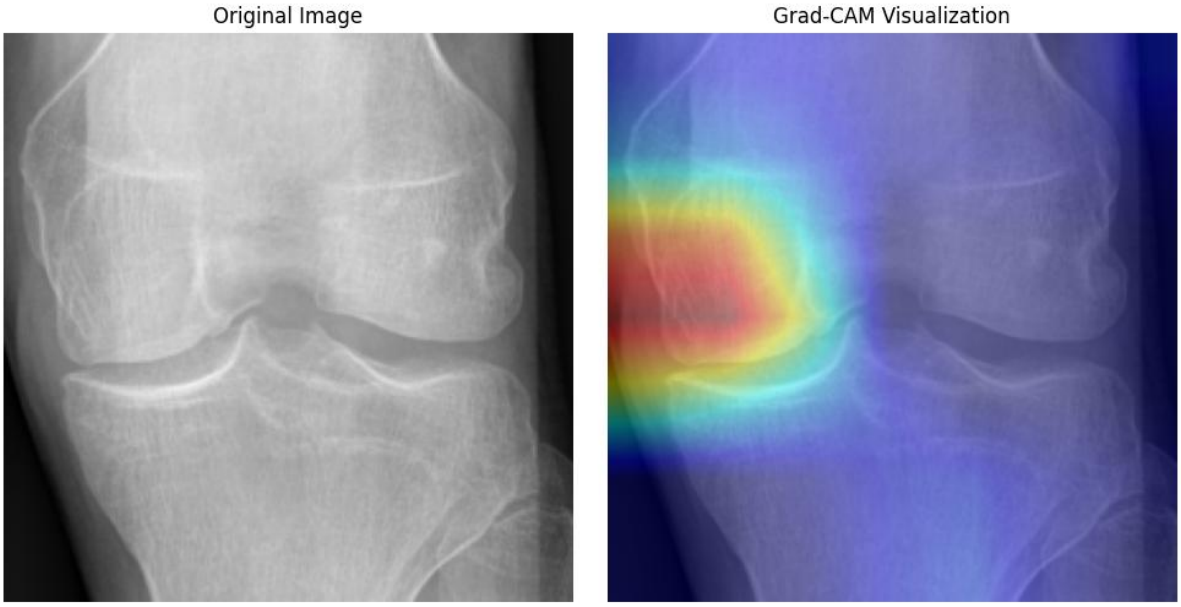


Figure 15 Grad-CAM Visualisation of a Knee X-ray Image

In the figure 15 above, the left panel displays the original knee radiograph, while the right panel presents the Grad-CAM overlay. The red-to-yellow regions in the Grad-CAM output indicate areas with the highest activation contributing to the model's classification. Notably, the highlighted region corresponds to the medial tibiofemoral joint, a common site of degeneration in knee osteoarthritis. This localisation suggests that the model has learned to associate structural narrowing and deformity in this area with advanced disease severity.

This form of visualisation acts as a qualitative confirmation that the model focuses on clinically relevant areas, reducing the risk of “black-box” decision-making. Such interpretability is especially vital in healthcare applications, where incorrect or non-transparent predictions may have significant consequences (Selvaraju et al., 2017).

In our study, Grad-CAM outputs were particularly useful during the validation phase to ensure that the model was not fixating on irrelevant image artefacts or margins. The consistent attention to pathological regions across multiple samples demonstrates the clinical alignment of the model’s internal representations.

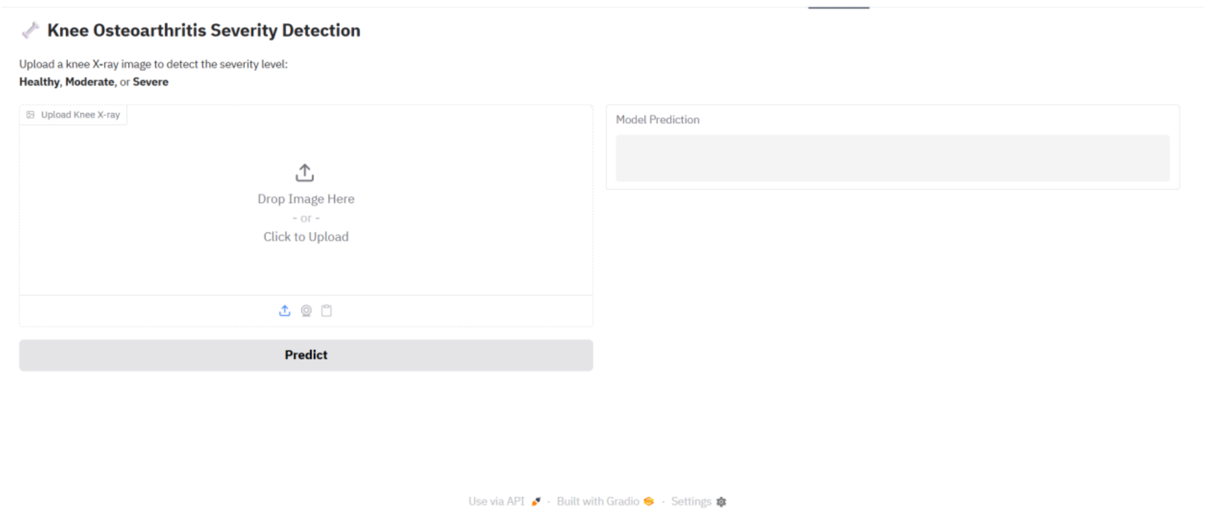
Furthermore, this approach aligns with best practices in medical AI, where model explainability is a key criterion for deployment in real-world settings (Ardila et al., 2019). It also aids radiologists in confirming the model's findings or identifying instances where human review is necessary, thereby enhancing collaborative diagnosis.

6.5 Web Application Deployment

Deploying the knee osteoarthritis (KOA) severity detection model as a web application was a critical step toward making the solution usable in real-world clinical environments. To bridge the gap between the trained deep learning model and its end-users, a lightweight yet functional interface was developed using the Gradio Python library. Gradio allows rapid prototyping and seamless deployment of machine learning models via an interactive web interface, eliminating the need for specialized software installation or programming expertise (Abid et al., 2019).

The chosen model for deployment was EfficientNetB3, which outperformed ResNet50 in both accuracy and generalizability (as demonstrated in previous sections). The deployed interface enables users—such as clinicians, radiology technicians, or even researchers—to simply upload a knee X-ray image and obtain a classification result indicating whether the image is classified as *Healthy*, *Moderate*, or *Severe*. This web-based approach supports real-time inference and is particularly beneficial for use in rural or resource-limited settings, where high-performance computing resources or skilled radiologists may not be readily available (Esteva et al., 2021).

Figure 16 below shows the initial interface of the Gradio web app. Users are prompted to either click or drag-and-drop a knee X-ray image into the upload area. This triggers the backend process, where the image is preprocessed (resized, normalized, and batched) and passed through the EfficientNetB3 model hosted on the same environment. The model processes the image and returns a severity prediction along with a confidence score ranging from 0 to 1, indicating the certainty of the classification.

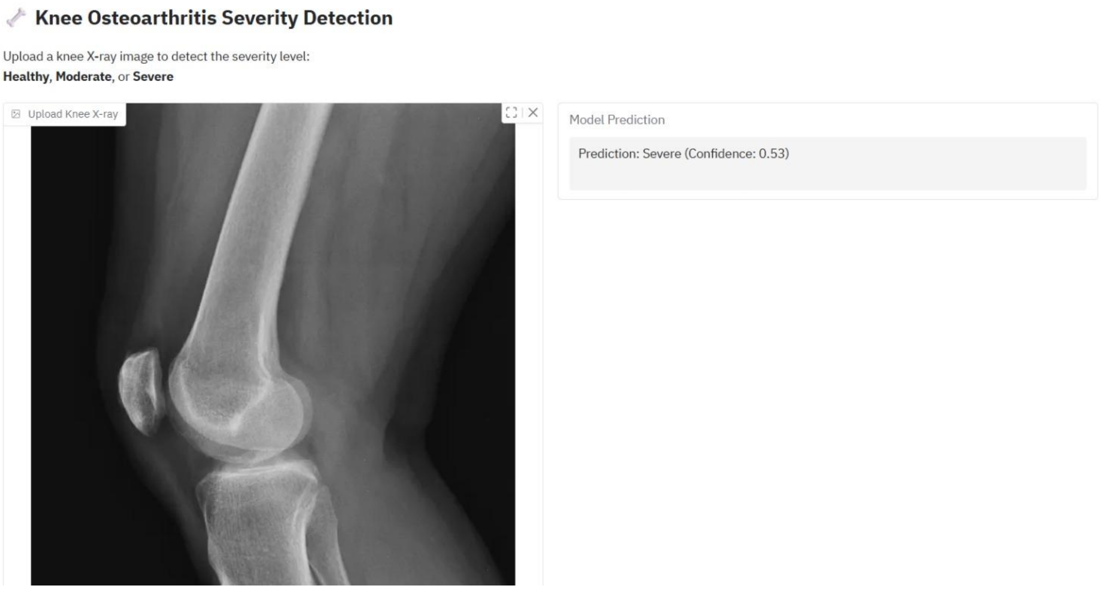


Knee Osteoarthritis Severity Detection

Upload a knee X-ray image to detect the severity level:
Healthy, **Moderate**, or **Severe**

Figure 16 Initial Interface of the Deployed Web Application for Knee Osteoarthritis Severity Detection.

In Figure 17, the result after an image upload is displayed. The application provides instant feedback by showing the predicted class (*e.g., Severe*) and a confidence value (*e.g., 0.53*). This form of result presentation not only enhances transparency but also supports decision-making in ambiguous cases, where borderline severity grades may lead to varied interpretations by human observers. According to Rajpurkar et al. (2017), such decision support tools can significantly reduce inter-observer variability and diagnostic delay in radiological workflows.



Knee Osteoarthritis Severity Detection

Upload a knee X-ray image to detect the severity level:
Healthy, **Moderate**, or **Severe**

Model Prediction

Prediction: Severe (Confidence: 0.53)

Figure 17 Web Application Output Showing Predicted Severity Class with Confidence Score after Uploading a Knee X-ray.

From a deployment standpoint, the use of Gradio provides several advantages:

- Accessibility: It is hosted in-browser and does not require local model execution.
- Efficiency: The interface is minimal yet functional, supporting quick predictions.
- Extensibility: It allows for future integration with Grad-CAM visualizations, electronic health records (EHRs), or additional patient metadata.

Furthermore, this deployment reflects the growing trend of integrating AI tools in telemedicine and clinical diagnostics, which is increasingly recognized as a solution to overcome disparities in healthcare delivery (Topol, 2019). This application serves as an early prototype demonstrating how AI-assisted tools can support orthopedic assessments without the need for deep AI expertise among end users.

Chapter 7 – Conclusion and Future Work

7.1 Conclusion

This research successfully explored the potential of deep learning models for automating the severity classification of knee osteoarthritis (KOA) using X-ray images. By focusing on two advanced convolutional neural network (CNN) architectures—EfficientNetB3 and ResNet50—the study demonstrated the application of transfer learning for medical image analysis, with practical implications for diagnostic efficiency in clinical environments.

The methodology encompassed rigorous steps: from data acquisition, preprocessing, and augmentation, to model training, evaluation, and deployment. A curated dataset of over 9,800 anonymised knee X-rays sourced from Kaggle served as the foundation, representing the three severity classes of KOA based on the Kellgren–Lawrence grading system. Extensive preprocessing ensured image uniformity, and targeted augmentation addressed class imbalance, particularly the underrepresentation of severe cases—an often-encountered challenge in medical datasets (Litjens et al., 2017; Shorten & Khoshgoftaar, 2019).

The results indicated that both models performed well in multi-class classification; however, EfficientNetB3 outperformed ResNet50 in several key performance metrics such as accuracy, precision, recall, and F1-score. This superiority can be attributed to EfficientNetB3's compound scaling technique, which optimally balances network depth, width, and resolution (Tan & Le, 2019). Visual explanations using Grad-CAM further validated the model's attention to diagnostically significant regions, enhancing interpretability and clinical trust (Selvaraju et al., 2017).

In addition to offline evaluation, the best-performing model—EfficientNetB3—was integrated into a user-friendly web application, enabling practical access for healthcare professionals and researchers. This end-to-end deployment not only illustrates technical feasibility but also lays the groundwork for future real-world adoption of AI tools in musculoskeletal radiology (Topol, 2019; Esteva et al., 2019).

Ultimately, the study contributes to the growing body of literature demonstrating the viability of AI-assisted diagnostic tools in healthcare. It reinforces the importance of balancing technical sophistication with ethical, social, and professional considerations in the design and deployment of AI systems.

7.2 Future Work

While the project accomplished its key objectives, several opportunities exist for future enhancement and research extension. One immediate direction would be to explore more advanced or specialised deep learning architectures tailored to medical imaging, such as Vision Transformers (Dosovitskiy et al., 2020), DenseNet (Huang et al., 2017), or ensemble models that combine predictions from multiple networks to increase robustness and generalisability.

Another potential avenue involves the incorporation of clinical metadata—such as patient age, gender, BMI, or history of joint pain—into the model pipeline. Multimodal learning frameworks that integrate radiographic data with structured clinical variables have been shown to improve diagnostic performance and enable personalised risk stratification (Miotto et al., 2017; Zhang et al., 2021).

Future iterations of this work could also be validated on larger, more diverse datasets sourced from real hospital archives or through collaborations with healthcare institutions. Cross-institutional validation would ensure better generalisation across imaging modalities, scanner settings, and population demographics—factors often neglected in isolated dataset studies.

In terms of deployment, integrating the system into electronic health records (EHRs) and clinical decision support systems (CDSS) would enhance its utility in real-time diagnostic workflows. Ensuring interoperability with HL7/FHIR standards and obtaining regulatory approvals (e.g., MHRA in the UK or FDA in the US) are key steps toward clinical translation (McKinney et al., 2020).

Lastly, expanding the system's functionality to detect early KOA signs or distinguish between other joint diseases (e.g., rheumatoid arthritis) could broaden its scope and relevance in orthopedic diagnostics. Such advancements, combined with continual retraining on updated datasets, would help maintain clinical relevance and effectiveness in evolving healthcare contexts.

REFERENCES

Abid, A., Farooqi, M., Zou, J. and Zou, J., 2019. Persistent anti-muslim bias in large language models. *arXiv preprint*, arXiv:2101.05783. Available at: https://arxiv.org/abs/2101.05783

Antony, B., Jones, G., Jin, X., Ding, C. and Cicuttini, F., 2017. Do knee abnormalities visualised on MRI explain knee pain in knee osteoarthritis? A systematic review. *British Journal of Sports Medicine*, 51(8), pp.700–708. Available at: https://bjsm.bmj.com/content/51/8/700

Apostolopoulos, I.D. and Mpesiana, T.A., 2020. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Physical and Engineering Sciences in Medicine*, 43(2), pp.635–640. Available at: https://link.springer.com/article/10.1007/s13246-020-00865-4

Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G. and Naidich, D.P., 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, 25(6), pp.954–961. Available at: https://www.nature.com/articles/s41591-019-0447-x

Bai, W., Suzuki, H., Huang, J., Francis, C., Tarroni, G., Oktay, O., Matthew, J., Rajchl, M. and Rueckert, D., 2021. Self-supervised learning for cardiac MR image segmentation by anatomical position prediction. *Medical Image Analysis*, 69, p.101977. Available at: https://www.sciencedirect.com/science/article/pii/S1361841520302097

Chen, J., Tang, Y., Miao, S. and Luo, J., 2020. Deep learning for medical image analysis. *Frontiers of Medicine*, 14(4), pp.417–430. Available at: https://link.springer.com/article/10.1007/s11684-020-0758-1

Chicco, D. and Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), pp.1–13. Available at: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6413-7

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp.248–255. IEEE. Available at: https://ieeexplore.ieee.org/document/5206848

Dosovitskiy, A., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint*, arXiv:2010.11929. Available at: https://arxiv.org/abs/2010.11929

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), pp.115–118. Available at: https://www.nature.com/articles/nature21056

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S. and Dean, J., 2019. A guide to deep learning in healthcare. *Nature Medicine*, 25(1), pp.24–29. Available at: https://www.nature.com/articles/s41591-018-0316-z

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F. and Schafer, B., 2018. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), pp.689–707. Available at: https://link.springer.com/article/10.1007/s11023-018-9482-5

Gulshan, V., et al., 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22), pp.2402–2410. Available at: https://jamanetwork.com/journals/jama/fullarticle/2588763

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.770–778. Available at: https://ieeexplore.ieee.org/document/7780459

Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.4700–4708. Available at: https://ieeexplore.ieee.org/document/8099726

Jaiswal, A., Gianchandani, N., Singh, D., Kumar, V. and Kaur, M., 2020. Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning. *Journal of Biomolecular Structure and Dynamics*, 40(4), pp.1945–1955. Available at: https://doi.org/10.1080/07391102.2020.1788642

Kermany, D.S., et al., 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5), pp.1122–1131.

Kingma, D.P. and Ba, J., 2015. Adam: A method for stochastic optimization. *arXiv preprint*, arXiv:1412.6980.

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. *Nature*, 521(7553), pp.436–444.

Litjens, G., et al., 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, pp.60–88.

McKee, M., 2020. Ethical issues in AI in health care: The long and winding road. *The Lancet Digital Health*, 2(4), pp.e158–e159.

McKinney, S.M., et al., 2020. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), pp.89–94.

Miotto, R., Wang, F., Wang, S., Jiang, X. and Dudley, J.T., 2017. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6), pp.1236–1246.

NHS England, 2019. The NHS Long Term Plan. [online] Available at: https://www.longtermplan.nhs.uk/

Oakden-Rayner, L., 2020. Exploring large-scale public medical image datasets. *Academic Radiology*, 27(1), pp.106–112. Available at: https://doi.org/10.1016/j.acra.2019.10.006

Pedregosa, F., et al., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, pp.2825–2830. Available at: https://doi.org/10.5555/1953048.2078195

Rajpurkar, P., et al., 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint*, arXiv:1711.05225. Available at: https://arxiv.org/abs/1711.05225

Raghu, M., Zhang, C., Kleinberg, J. and Bengio, S., 2019. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*, 32.

Said, E.A., Salem, A.B.M., Ghoneim, A.S. and El-Fishawy, N.A., 2021. Deep learning techniques for diagnosing knee osteoarthritis from medical images. *Computer Methods and Programs in Biomedicine*, 198, p.105794.

Selvaraju, R.R., et al., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp.618–626.

Shankar, K., Perumal, E. and Tiwari, P., 2021. Comparative performance analysis of VGGNet, AlexNet, and GoogleNet for breast cancer classification. *Materials Today: Proceedings*, 45, pp.1145–1149.

Shorten, C. and Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), p.1.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1), pp.1929–1958.

Tan, M. and Le, Q., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pp.6105–6114.

Tiulpin, A., et al., 2018. Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. *Scientific Reports*, 8(1), p.1.

Topol, E.J., 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), pp.44–56.

World Health Organization, 2023. WHO Global Strategy on Digital Health 2020–2025. [online] Available at: https://www.who.int

Yosinski, J., Clune, J., Bengio, Y. and Lipson, H., 2014. How transferable are features in deep neural networks?. In *Advances in Neural Information Processing Systems*, pp.3320–3328.

Zhang, Z., Liu, Q., Wang, Y. and Zhang, B., 2021. A review on the development of medical image diagnosis using deep learning. *Journal of Healthcare Engineering*, 2021. Available at: https://doi.org/10.1155/2021/6656724