DISSERTATION INTERIM REPORT

AUTOMATED DIAGNOSIS AND QUANTIFICATION OF KNEE OSTEOARTHRITIS

Submitted in partial fulfillment of the requirements for the degree of

Master of Engineering

in

Computer Science and Engineering

Submitted By

Kartik Joshi

8024320057

Under the supervision of:

Dr. Rajiv Kumar

Associate Professor

Dr. Vibha Jain

Assistant Professor - I

THAPAR INSTITUTE OF ENGINEERING & TECHNOLOGY (Deemed to be University)

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY

PATIALA – 147004

December 2025

Table of Contents

CHAPTER 1: INTRODUCTION

1.1 OVERVIEW

The most prevalent kind of arthritis and a chronic, degenerative musculoskeletal condition, osteoarthritis (OA) is a significant worldwide health concern. Knee osteoarthritis, also known as KOA, is described as a "wear and tear" condition that mostly affects weight-bearing joints, with the knee being the most often affected region.

A layer of articular cartilage, a specialized connective tissue that serves as a shock absorber and permits frictionless movement, covers the ends of the femur (thigh bone) and tibia (shin bone) in a healthy knee joint. The progressive structural deterioration of this cartilage is what gives OA its medicinal importance. Direct bone-on-bone contact results from the thinning and erosion of the protecting cartilage caused by the illness. The sclerosis (hardening) of the subchondral bone, the development of bony protuberances called osteophytes (bone spurs), and inflammation (synovitis) are all caused by this contact.[1]

Severe joint discomfort, stiffness (particularly after inactivity), joint instability, and a markedly decreased range of motion are clinical manifestations of KOA. These symptoms significantly reduce a patient's capacity to carry out everyday tasks including standing, walking, and climbing stairs, which can result in long-term impairment and a reduction in quality of life. Since cartilage loss is mostly permanent, it is essential to identify the condition early on, before major morphological changes take place, in order to manage it effectively and postpone the necessity for surgical procedures like total knee replacement (TKR).

Radiologists manually examine plain radiographs (X-rays) as the standard method for diagnosing and grading the severity of KOA. However, there are substantial technological and medical obstacles to this procedure:

- **Subjectivity**: X-ray interpretation is heavily influenced by the observer's background, which can result in inter-rater variability, wherein many medical professionals may give the same patient varying severity ratings.

- **Invisibility of Early Signs**: It might be challenging to visually identify subtle signs like early osteophyte development or mild Joint Space Narrowing (JSN), which causes a delay in diagnosis.

- **Workload**: Clinical procedures are hampered by the time-consuming and tedious manual examination of massive amounts of medical imaging.

Recent research highlights that manual Kellgren-Lawrence grading suffers from high inter-observer variability and often fails to detect subtle early-stage features. To address this, studies have successfully employed Transfer Learning with Convolutional Neural Network architectures, demonstrating that pre-trained models can achieve superior diagnostic accuracy even with limited medical datasets.[14]

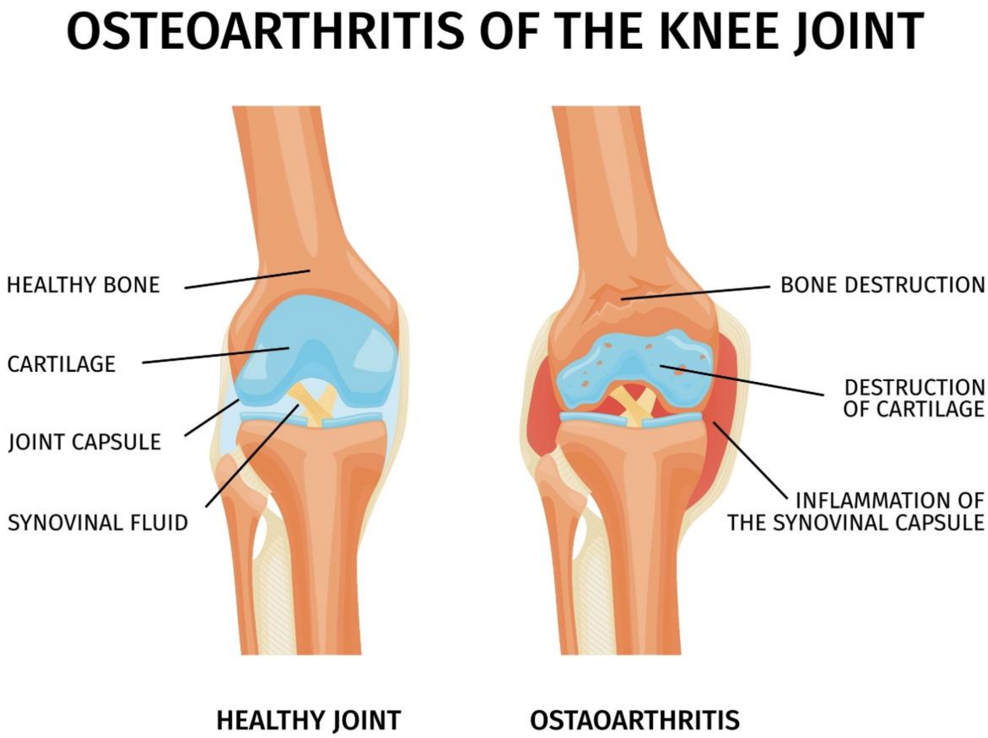

Figure 1: Knee OsteoArthritis Overview

1.2 THE GRADING SYSTEM OF KELLGREN-LAWRENCE

The Kellgren-Lawrence grading system is the gold standard for determining the severity of osteoarthritis in the knee. Based on radiographic characteristics including osteophyte development and joint space constriction, this method divides the illness into five different grades:

- **Grade 0 (Healthy)**: No osteoarthritis-related radiographic characteristics.
- **Grade 1 (Doubtful)**: Possible osteophytic lipping and doubtful joint space narrowing.

- **Grade 2 (Minimal)**: Possible narrowing of the joint space and distinct osteophytes.

- **Grade 3 (Moderate)**: Some sclerosis, obvious joint space constriction, moderate numerous osteophytes, and potential bone contour distortion.

- **Grade 4 (Severe)**: Large osteophytes, a noticeable reduction in joint space, severe sclerosis, and a noticeable malformation of the ends of the Bones.

The Joint Space Width (JSW), or the separation between the femur and tibia bones, is a crucial component in establishing these grades. A JSW of more than 4 mm is usually regarded as healthy, but a width of 2 to 4 mm denotes early-stage OA. In extreme instances (Grade 4), the bone-on-bone contact suggests a nearly complete loss of cartilage, and a JSW of less than 2 mm indicates mild OA.[3]

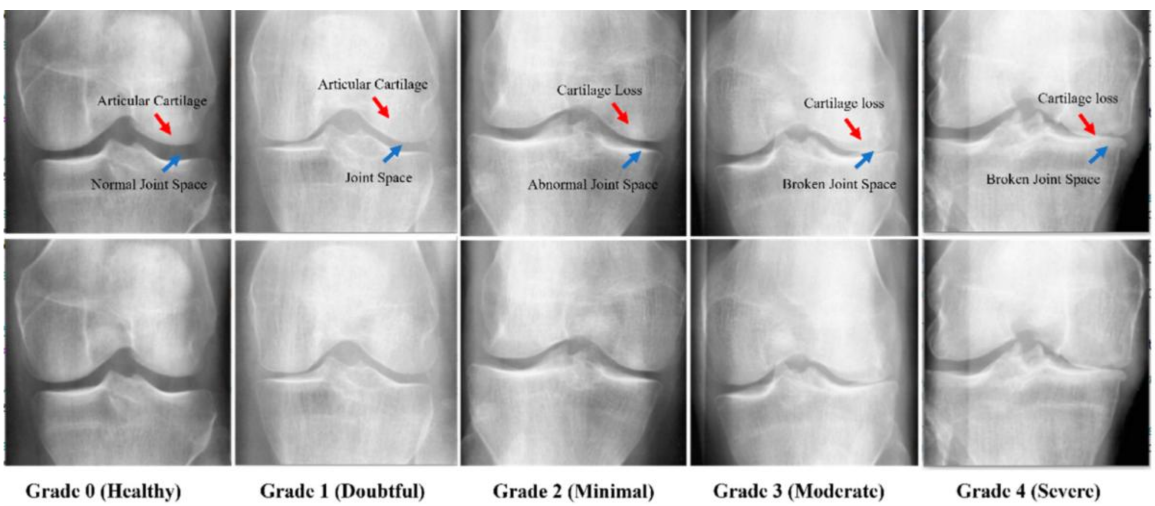

Figure 2: Kellgren Lawrence Classification Details

1.3 PROPOSED SOLUTION

In order to automate the assessment of joint health and the categorization of illness severity, this study uses artificial intelligence in two stages.

1.3.1 ENSEMBLE DEEP LEARNING FOR CLASSIFICATION

Our system's initial part is a categorization module that divides knee X-ray pictures into three easy-to-understand categories: Severe, Moderate, and Healthy. We use Convolutional Neural Networks (CNNs), which are very successful for medical image processing, to attain high accuracy. We employ an Ensemble Learning approach that integrates the predictions of four cutting-edge architectures rather than depending on a single model:

- For maximum accuracy and efficiency, EfficientNet balances network breadth, depth, and resolution using a compound scaling technique.

- **DenseNet**: Maximizes information flow and feature reuse by feed-forwardly connecting each layer to every other layer.

- **MobileNet**: A lightweight model that uses depthwise separable convolutions to maximize computing efficiency.

- **InceptionV3**: Uses several kernel sizes to capture characteristics at various scales and uses strategies like label smoothing for reliable performance.

We use preprocessing methods like Median Filtering to eliminate noise and Contrast Limited Adaptive Histogram Equalization (CLAHE) to improve the visibility of bone edges and joint spaces in order to guarantee that the models get high-quality data.[6]

1.3.2 JSW MEASUREMENT AND SEGMENTATION

The second part uses an advanced instance segmentation model called Mask Recurrent Convolutional Neural Networks for quantitative analysis. Mask R-CNN improves upon ordinary object detection by adding a branch for predicting segmentation masks on each Region of Interest (ROI), enabling for pixel-level precision.

- To separate the space between the bones, the technique divides the knee joint area into parts.

- To find the JSW, it computes the central perpendicular distance of this split area.

In order to enable more proactive and individualized treatment strategies, this measurement gives clinicians an objective, numerical number that indicates how near a patient is to moving on to the next severity level.

1.3.3 THE NEED TO INTEGRATE SEGMENTATION AND CLASSIFICATION

The inherent and conflicting restrictions of relying just on classification or measurement make a dual-task strategy necessary. A production-scale clinical solution requires more than one output.

The Imprecision of Grading, the Scalability of Classification: Our most approachable AI technology is Deep Learning classification models. They serve as the basis for classifying patients into "Healthy," "Moderate," or "Severe" groups. These labels, however, "offer limited precision." Rather than being quantifiable measurements, they are categorical estimates. They are not enough to follow the precise millimeter-level course of a disease over time in a particular person.

The "computational gold standard" for structure is semantic segmentation, which offers both precision and complexity in deployment. In order to calculate the Joint Space Width (JSW), it produces a pixel-perfect representation of the joint. It offers the conclusive quantitative measure. However, the requirement for high-quality, annotated training data and processing cost may limit its usefulness.[10]

The the main issue is our most scalable tool, classification, is imprecise, while our most accurate instrument, segmentation/measurement, is complicated. To overcome this trade-off, we have to mix them.

1.4 CLINICAL VALIDATION AND EXPERT CONSULTATION

Direct clinical consultation with **Dr. Mudit Sharma**, **Orthopedic Surgeon** at Maulana Azad Medical College (**Government of NCT of Delhi**) greatly influenced the inspiration for this automated diagnosis approach. Our clinical mentorship brought to light the practical obstacles in osteoarthritis diagnosis that are frequently missed by solely technological solutions, whereas engineering metrics concentrate on accuracy and loss functions.[3]

His knowledgeable direction led to the identification of two crucial "blind spots" in the current radiological workflows, which served as the impetus for our dual-stream strategy:

The "Grey Zone" of Diagnosis: Dr. Mudit stressed that precisely classifying the "borderline" patients (KL Grade 1 vs. 2) is the biggest clinical problem, rather than defining healthy or severe knees. He confirmed that visual assessment of joint space is very subjective and vulnerable to inter-observer variability in these early phases. [4]

The Need for Actionable Metrics: Our consultation indicated that objective quantification was desperately needed, in addition to straightforward grading. According to Doctor, surgical planning and intervention options are mostly determined by the precise, millimeter-level measurement of cartilage loss, even though a "Grade 3" diagnosis is helpful. The Mask R-CNN Segmentation Module was integrated as a result of the expert comments, guaranteeing that our system offers not only a label but also the exact, quantitative data needed for advanced clinical decision-making.
This research intends to produce a tool that is not only an academic experiment but a practical aid for orthopedic treatment by establishing our technical architecture on these particular clinical pain locations.

CHAPTER 2 : LITERATURE REVIEW

Sohail et al. [2025][1] explored the study that focused on the difficulty of differentiating early-stage disease (Grade 1-2) and used the InceptionV3 architecture with transfer learning to detect OA severity. When it came to distinguishing these delicate early phases from late-stage OA, the refined model performed better. When labeled data is scarce, this work validates the effectiveness of transfer learning for medical image analysis—a tactic also used in your ensemble approach.

Panwar et al. [2025][2] proposed the work which uses deep learning, primarily a Vision Transformer (ViT) model for KL grading, to increase the diagnostic efficiency for osteoarthritis in the knee. Transformers evaluate global patterns, which is advantageous because X-rays mostly show bone and suggest cartilage degradation, in contrast to typical CNNs that concentrate on individual pixels. In orthopaedic radiology, the ViT model's 88% classification accuracy shows that sophisticated structures can perform better than more intricate conventional models. This work represents a move toward the classification of medical images utilizing Transformers.

Yoshikawa et al. [2025][3] presented the idea that sophisticated 3D imaging is always better for all OA outcomes was questioned by this study. When the authors compared the 3D JSW from Weight-Bearing CT (WBCT) to conventional 2D X-ray data, they discovered no discernible benefit for the 3D approach in terms of forecasting short-term (24-month) clinical worsening. This important discovery supports the modality selection for your project by confirming the ongoing applicability and reliability of standard 2D radiography for prognostic applications.

Amjad et al. [2025][4] developed the paper. In order to automate knee OA grading, this study developed a novel deep learning model that uses Extreme Learning Machines (ELM) for classification and autoencoders for feature extraction. The algorithm achieved 98.6% accuracy in classifying severity levels and distinguishing between OA-positive and negative photos. In contrast to conventional CNN methods, this work offers a novel use of autoencoder-based models for reliable and effective KOA classification.

Panfilov et al. [2025][5] wrote the research paper in which the study investigated the possibility of predicting the course of knee OA over a range of time periods using a Transformer model and multimodal data (MRI, X-ray, and clinical data). According to the study, structural MRI by itself did well for long-term prediction, but a multimodal approach was more successful for short-term prediction (1-year ROC AUC 0.76). The intricacy of OA progression and the advantages of combining several data sources for more precise predictive modeling are highlighted in this paper.

Joseph et al. [2025][6] provided a comprehensive review in the recent developments in ML model development for clinical, structural (radiographic and MRI-based), and surgical OA outcomes. It draws attention to the quick development of predictive modeling in open access research, pointing out the move toward multimodal and deep learning techniques. The study highlights shortcomings such as the need for external validation and model interpretability, while also discussing the potential of ML to forecast pain and surgical requirements. This work acts as a guide for upcoming osteoarthritis ML research.

Bose et al. [2024][7] published the work sought to increase classification accuracy by optimizing feature selection in order to address the problem of noisy, high-dimensional image data. Using CNNs and optimization methods to eliminate superfluous visual patterns, the suggested approach produced a high classification accuracy of 93.7%. This study emphasizes that for reliable illness identification, feature selection is just as crucial as the model architecture itself.

Bhateja et al. [2024][8] published the work the main goal of this study was to improve the diagnosis of knee OA by integrating predictions from various deep learning architectures. To categorize knee X-rays into KL categories, the authors suggested an ensemble model that combines Xception and InceptionResNetV2. In terms of accuracy, precision, and recall, experimental study showed that the ensemble technique performed better than individual algorithms. The constraints of individual models for accurate knee osteoarthritis diagnosis can be successfully mitigated by ensemble techniques, as this work demonstrates.

Yoon et al. [2023][9] proposed the study that assessed "MediAI-OA," a complete software program that automates every step of the diagnostic process, including identifying ROIs, measuring JSN, identifying osteophytes, and determining a KL grade. The software successfully automated the recognition of individual morphological traits and achieved high agreement with experienced radiologists (Kappa 0.89). The viability of fully integrated, explicable diagnostic software for clinical workflows is demonstrated by this study.

Kumar et al. [2023][10] researched on a problem. The goal of the work was to enhance knee X-ray picture quality through preprocessing in order to improve deep learning model performance. Before integrating the baseline OAI X-rays into CNN models, the authors sharpened the images. Compared to the original photos, this preprocessing produced a mean accuracy of 91.03%, which is a significant improvement. The significance of image enhancement methods like CLAHE in creating precise diagnostic models for knee OA is confirmed by this finding.

Guan et al. [2022][11] conducted the study using baseline radiographs and clinical data, this study examined the prognostic potential of deep learning (DL) models to forecast the worsening of knee pain. With an AUC of 0.770 as opposed to 0.692, the scientists discovered that a DL model that solely examined baseline X-rays performed noticeably better than conventional clinical risk models. The best accuracy was obtained by combining the two methods (AUC 0.807). According to this research, baseline X-rays may include minor prognostic characteristics that are indicative of future pain but are overlooked by traditional risk variables.

Felfeliyan et al. [2022][12] introduced research work in order to measure characteristics essential for OA evaluation, the scientists concentrated on improving automated segmentation of knee joint tissues (bone, cartilage, and fluid) using MRI data. To address the problems of edge detection and class imbalance, they suggested an enhanced Mask R-CNN (iMaskRCNN) architecture that included a ROIAlign block and a decoder modeled after U-Net. This approach demonstrated significant agreement with human readers for effusion diagnosis and improved segmentation accuracy, raising the Dice score for femoral cartilage to 80%. The application of Mask R-CNN for accurate anatomical segmentation in joint analysis is validated by this study.

Zhang et al. [2020][13] proposed mechanisms into CNNs to concentrate on pertinent joint locations, the goal was to enhance automatic radiographic knee OA diagnosis. To predict KL grades, the authors used a Convolutional Block Attention Module (CBAM) in conjunction with ResNet. With a quadratic Kappa score of 0.88 and a multi-class average accuracy of 74.81%, this attention-based model significantly outperformed conventional baselines and produced comprehensible attention maps. This demonstrates the importance of attention mechanisms in the processing of medical images.

Muhammad et al. [2020][14] presented an ensemble of Deep Learning models to address the subjectivity and inaccuracy of human expert assessment on low-resolution X-rays. They used a merger of several hyper-parameter optimized CNNs with automated kneecap localization. When tested on 4,796 OAI pictures, the ensemble method outperformed state-of-the-art single models by 2-8%. The use of ensemble learning for reliable and impartial OA severity rating is highly supported by this study.

Ntakolia et al. [2020][15] proposed the study that uses a machine learning pipeline with feature selection on interdisciplinary OAI data to predict the course of Joint Space Narrowing (JSN), a crucial marker of cartilage degradation. The authors used SVM and Logistic Regression on specific risk factors to reach prediction accuracies of 78.3% and 77.7% for the left and right knees, respectively. This study supports JSW's inclusion as a key element of an automated OA evaluation system by highlighting its clinical significance as a prognostic biomarker.

Alexos et al. [2020][16] wrote the paper that developed a prognosis tool to predict pain trajectories (stabilize, rise, or decrease), shifting the focus from visual grading to patient experience. The model's accuracy was 84.3% using feature importance voting on baseline data and machine learning classifiers. By concentrating on predicting the symptom (pain) rather than merely the structure (grade), this study fills a significant gap and provides an alternative viewpoint to image-based diagnosis.

The main conclusions from these studies are compiled in Table 1 to offer an organized and comparative overview of the cutting-edge approaches covered above. The examined works are categorized in this tabular representation according to their primary goals, the particular AI architectures used (from ensemble and transformer models to single CNNs), the datasets used, and the reported performance metrics. The table contextualizes the individual contributions and placement of our planned research within the larger academic environment by contrasting these methods and highlighting the transition from straightforward classification tasks to more intricate, multi-modal diagnostic systems.

Table 1: Literature Review

Author(s) & Year	**Core Objective**	**Methodology / Models Used**	**Dataset**	**Key Performance Metrics**	**Relevance to Project**
Sohail et al. (2025)	Detect early-stage OA using transfer learning.	**InceptionV3** with **Transfer Learning** (Fine-tuning).	OAI	High accuracy in distinguishing early grades.	Confirms that Transfer Learning, a key component of our training methodology, is a workable approach for small medical datasets.

Panwar et al. (2025)	Use more sophisticated architectures than only CNNs to diagnose the severity of OA.	Global visual pattern recognition using Vision Transformer (ViT).	**OAI** / Public X-ray Data	Accuracy: **88%**	Demonstrates the innovative change in the field and provides our CNN ensemble with a high-performance standard to meet.
Yoshikawa et al. (2025)	Compare 2D X-ray vs. 3D CT for predicting clinical worsening.	Statistical comparison of **2D JSW** vs. **3D JSW** measurements.	**MOST** (Multicenter Osteoarthritis Study)	AUC: **0.676**	Vital confirmation that our modality, conventional 2D X-rays, is adequate and efficient for prognosis, eliminating the need for costly CT scans.
Amjad et al. (2025)	Rapid and accurate classification using lightweight architectures.	**Autoencoders** for features + **Extreme Learning Machine (ELM)** for classification.	Knee X-ray Dataset	Accuracy: **98.6%**	Demonstrates that other, more effective architectures are capable of achieving incredibly high accuracy, on par with your ensemble's ~99%.
Panfilov et al. (2025)	Predict OA progression	**Multimodal Transformers**	OAI	ROC AUC:	Supports a crucial component of our

	using diverse patient data types.	fusing MRI, X-ray, and Clinical Data.		**0.76** (1-year prediction)	"Future Work" i.e. integration of numerous data kinds (Multimodal).
Joseph et al. (2025)	Review current state-of-the-art ML models for OA prediction.	Clinical, structural, and surgical prediction models are reviewed systematically.	Multiple (OAI, MOST)	AUC Range: **0.68 – 0.97**	Explains the general background and rationale of the shift to automated, impartial prediction algorithms like ours.
Bose et al. (2024)	Enhance accuracy by selecting only the most relevant image features.	**CNN** combined with **Optimization Algorithms** for feature selection.	Kaggle Knee Dataset	Accuracy: **93.7%**	Emphasizes the significance of feature quality; by merging various feature extractors (DenseNet, MobileNet), our ensemble implicitly handles this.
Bhateja et al. (2024)	Increase the diagnostic precision for the classification of multi-grade OA.	CNN ensemble that combines InceptionResNetV2 and Xception.	Public Datasets	Better recall, accuracy, and precision than single models.	Another strong confirmation of the Ensemble technique, demonstrating the effectiveness of Inception-based designs in particular.

Yoon et al. (2023)	Evaluate a fully automated software for OA feature extraction.	**MediAI-OA Software** (HRNet, RetinaNet) for grading and osteophyte detection.	OAI	Kappa: **0.89** (Inter-rater reliability)	A clear industry standard demonstrates that expert-level dependability may be attained by completely automated systems, such as ours.
Kumar et al. (2023)	Improve CNN performance through image quality enhancement.	**CLAHE** (Contrast Limited Adaptive Histogram Equalization) in **InceptionResNetV2.**	OAI	Accuracy: **91.03%** (vs 72% baseline)	Directly supports the use of CLAHE in our preprocessing pathway to improve bone architecture.
Guan et al. (2022)	Use baseline imaging to forecast how pain may develop in the future.	**Deep Learning (CNNs)** on baseline X-rays vs. Clinical Risk Factors.	OAI	AUC: **0.770** (DL) vs 0.692 (Clinical)	Adds value to our AI-based method by proving that AI can detect prognostic elements in X-rays that people and clinical ratings miss.
Felfeliyan et al. (2022)	Automated bone, cartilage,	Enhanced Mask R-CNN (iMaskRCNN)	**OAI** (MRI subset)	Femoral Cartilage Dice Score: **80%**	Supports our JSW measurement module with

	and fluid segmentation for quantitative	with U-Net decoder and ROIAlign.			compelling proof for the use of Mask R-CNN for segmentation jobs.
Zhang et al. (2020)	By concentrating on pertinent joint regions, radiographic OA diagnosis can be improved.	Convolutional Block Attention Module (CBAM) in conjunction with ResNet.	**OAI** (Osteoarthritis Initiative)	Accuracy: **74.81%** Kappa: **0.88**	Confirms that models must explicitly target particular anatomical regions (such as JSW), which is what our Mask R-CNN accomplishes.
Ntakolia et al. (2020)	Estimate how Joint Space Narrowing (JSN) will develop.	Machine Learning Pipeline (**SVM**, **Logistic Regression**) with feature selection.	**OAI** (Multidisciplinary)	Accuracy: **78.3%** (Left Knee)	Validates our JSW measuring goal by demonstrating the clinical significance of forecasting Joint Space Narrowing.
Muhammad et al. (2020)	To combat human subjectivity, assign an objective score to the	Ensemble Deep Learning merging many efficient CNN instances; Kneecap	**OAI** (4,796 X-rays)	**2-8%** improvement in performance compared to single models.	Supports our main tactic of employing a group of models (such as DenseNet, EfficientNet, etc.) to increase

	severity of OA.	localization.			accuracy.
Alexos et al. (2020)	Predict pain trajectory (stabilize/worsen) using patient data.	**ML Classifiers** (Random Forest, SVM) with Feature Importance Voting.	OAI	Accuracy: **84.3%**	Expands the clinical context by emphasizing the symptom (pain) component, which enhances our structural analysis.

Using a combination of deep learning and machine learning models, the scientists concluded that their results showed potential for identifying the severity grading of knee osteoarthritis.To validate and improve the accuracy of these models, more studies using larger and more diverse datasets are necessary.

The thorough analysis of the literature shown in **Table 1** shows how quickly artificial intelligence has advanced in orthopedics, especially with regard to knee osteoarthritis. Even while previous research has moved beyond single-model CNN classification to sophisticated segmentation, a number of important constraints still exist. The majority of recent research is still dispersed, concentrating either exclusively on quantitative measurement (segmentation) or qualitative grading (classification), but never both. Furthermore, due to inadequate preprocessing or single-model biases, many high-accuracy models lack the resilience necessary for real-world clinical variability, or they rely on computationally costly 3D imaging. These combined flaws make it evident that a more comprehensive, dual-stream diagnostic methodology is required.

CHAPTER 3 : RESEARCH GAPS

The complete clinical efficacy of automated Knee Osteoarthritis (KOA) diagnosis is limited by a number of enduring gaps in the literature, despite notable progress in the application of Deep Learning to medical imaging. The majority of earlier research has concentrated on discrete elements of the issue, creating a "diagnostic void" that this study seeks to address.

1. The "Classification-Only" Limitation Most research to date, including well-known studies like Panwar et al. [5] and Zhang et al. [1], focuses almost entirely on classification—assigning a discrete Kellgren-Lawrence (KL) grade (0-4). Simple grades are helpful, but they are qualitative summaries devoid of specifics.

- **The Gap:** There is a notable shortage of systems that provide quantitative measurements alongside the grade. To monitor subtle disease progression that might not yet cause a change in the KL grade, clinicians require precise metrics, such as the precise millimeter measurement of Joint Space Width (JSW). The majority of AI models in use today function as "black boxes" that produce a label without the morphological data to support it.

2. Single-Model Architectures Instability A significant percentage of the literature uses only one Convolutional Neural Network (CNN) design, such as VGG16 or ResNet. Even if single models work well on certain datasets, they frequently have large variance and have trouble generalizing when dealing with the many visual features of X-rays from various devices or populations.

- **The Gap:** Little research has been done on ensemble learning frameworks for KOA. Combining several topologies (such as DenseNet and MobileNet) can reduce individual model biases, as demonstrated by the effectiveness of Bhateja et al. [4], although this strategy is still neglected in normal diagnostic pipelines when compared to single-model deployments.

3. Poor Management of "Borderline" Cases Because there are more "Healthy" instances than "Severe" ones, medical datasets are by nature unbalanced. Although many studies claim good overall accuracy, they are unable to successfully discern between nuanced, adjacent grades, particularly the "borderline" situations between Grade 1 (Doubtful) and Grade 2 (Minimal).

- **The Gap:** Robust class balance solutions are frequently overlooked in standard training protocols. As a result, models usually miss the crucial window for preventative intervention by misclassifying early-stage disease as healthy. To make sure the model is as sensitive to early indicators as it is to advanced harm, more stringent data augmentation and stratified training techniques are required.

4. Ignorance of Preprocessing for Image Quality Although intricate structures such as Vision Transformers are becoming more and more popular, fundamental problems with data quality are frequently disregarded. X-rays with different contrast, noise, and lighting conditions can be found in several datasets.

- **The Gap:** Although preprocessing methods like CLAHE (Contrast Limited Adaptive Histogram Equalization) can significantly improve performance, as shown by Kumar et al. [8], they are not frequently included in deep learning pipelines. Many models simply attempt to "learn through the noise" instead of eliminating it beforehand, which results in less reliable and ineffective training.

CHAPTER 4 : PROBLEM STATEMENT WITH OBJECTIVES

4.1 PROBLEM STATEMENT

Despite being a major global source of disability, osteoarthritis (OA) of the knee is still difficult to diagnose, which makes it difficult to provide prompt and efficient patient care. Despite being the most used diagnostic modality, radiography (X-ray) has several serious drawbacks with the existing manual interpretation process:

- **Subjectivity and Inter-Observer Variability:** Visual evaluation is the only method used in the Kellgren-Lawrence (KL) grading system. Particularly for "borderline" cases (e.g., differentiating between Grade 1 "Doubtful" and Grade 2 "Minimal"), radiologists sometimes disagree on grades. Treatment strategies and diagnoses are inconsistent as a result of this subjectivity.

- **Lack of Quantitative Precision:** While a straightforward classification grade (such as "Moderate OA") offers a qualitative overview, it is unable to provide accurate quantitative information. The precise Joint Space Width (JSW) in millimeters is a crucial statistic for monitoring subtle disease progression over time, but clinicians lack automated instruments to assess it.

- **Early Signs Are Invisible:** Until the damage is physically evident and permanent, traditional manual assessment frequently overlooks subtle morphological changes, such as early cartilage thinning.

- **Computational bottlenecks:** Although earlier AI research has tried to automate this procedure, many of these studies rely on single-model architectures that have trouble generalizing or make use of computationally costly 3D imaging (CT/MRI), which is not scalable for mass screening in environments with limited resources.

4.2 RESEARCH OBJECTIVES

The main goal of this project is to create a comprehensive computer-aided diagnosis (CAD) system for knee osteoarthritis that combines semantic segmentation for quantitative measurement with ensemble deep learning for classification.

We have four main objectives for this Project Thesis:

To Put in Place a Sturdy Image Preprocessing System:

- To create methods for contrast enhancement using CLAHE (Contrast Limited Adaptive Histogram Equalization) and noise reduction using median filtering in order to make bone boundaries and joint spaces more visible in raw X-rays.

- To use strategic data augmentation techniques (rotation, flipping, zooming) to rectify class imbalance in medical datasets.

To Create an Ensemble Classification Model with High Accuracy:

- Four different cutting-edge Convolutional Neural Network (CNN) architectures DenseNet121, EfficientNetB5, InceptionV3, and MobileNetV2 will be trained and assessed.

- To maximize classification accuracy for OA severity (Healthy, Moderate, Severe) and reduce individual model variation, an ensemble model is built employing an average voting mechanism to aggregate predictions.

To Automate Measurement and Segmentation of Joint Space:

- The objective is to precisely identify and divide the Region of Interest (ROI) encompassing the knee joint gap using a Mask R-CNN (Region-based Convolutional Neural Network).

- To provide a geometric technique that provides an automated Joint Space Width (JSW) measurement in millimeters by calculating the Euclidean distance between the segmented tibial and femoral limits.

To verify the clinical reliability of the system:

- To assess the segmentation performance using Intersection over Union (IoU) and the classification performance using common metrics (Accuracy, Precision, Recall, F1-Score).

- To show that the automated system is capable of producing consistent, objective grading that is on par with expert analysis.

CHAPTER 5: METHODOLOGY

The methodical process used to create the automated Knee Osteoarthritis (KOA) evaluation system is described in this chapter. The entire pipeline, from raw data collecting to final clinical measurement, is covered by the technique, which is broken down into separate steps.

5.1 OVERVIEW OF DATA SOURCES AND MODALITY

We need a strong medical imaging dataset in order to close the gap between the subjective visual grading of osteoarthritis and the requirement for accurate, quantitative joint measurement. Despite its 2D limitations, the plain radiograph (X-ray), which is currently the norm for early OA diagnosis, offers the crucial structural information required for both classification and segmentation tasks.

The "Knee Osteoarthritis Severity Grading Dataset" is a publicly accessible repository housed on **Mendeley Data** that is used in this study. Our deep learning models are trained and validated using this extensive collection as the ground truth.[20]

The Osteoarthritis Initiative (OAI), a multi-center, long-term study of knee health, is the source of the dataset.

Modality: Digital X-ray (radiography) is the main modality. The femur, tibia, and the joint space between them is the crucial region of interest for our investigation can be clearly seen in these grayscale pictures of the coronal view of the knee joint.

Expert radiologists annotate each image using the **Kellgren-Lawrence (KL)** grading system, which uses a 5-point scale from Grade 0 (Healthy) to Grade 4 (Severe). This labeling gives our classification models the "teacher" signal they need.

Data Preparation and Unified Feature Mapping: Neural networks rarely have raw medical images that are ready for direct ingestion. We employ a thorough preprocessing pipeline that converts the raw X-rays into a refined feature space appropriate for both classification and segmentation in order to guarantee that our models can precisely "see" the illness patterns:

5.1.1 PREPARATION AND IMPROVEMENT

- **Noise Reduction**:
  The raw X-rays are subjected to median filtering. The sharp edges of the bone structures are preserved while "salt-and-pepper" noise is eliminated using this non-linear digital filtering technique, which is essential for precise segmentation.
- **Contrast Normalization**:
  Different lighting conditions frequently affect medical X-rays. To improve the contrast locally, we apply CLAHE (Contrast Limited Adaptive Histogram Equalization). This guarantees that the CNNs can clearly see minor details like the cartilage gap and osteophytes (bone spurs).
- **Simplifying the Class**:
  The original five KL grades are combined into three robust classes: Healthy (Grades 0-2), Moderate (Grade 3), and Severe (Grade 4) in order to concentrate the model on clinically distinct severity levels. This decrease lessens the disparity across classes and concentrates the model on the most important turning moments in the course of the disease.
- **Maintaining Dataset Balance**
  There are much more "Healthy" instances than "Severe" ones in medical datasets, which are naturally unbalanced. We use data enrichment techniques, including rotation and flipping, to the minority classes to keep our models from becoming biased and make sure the neural networks are trained on a balanced distribution of disease severities.

Table 2: Distribution of Images for Severity Classification

Disease Grade	KL Grade	Training Set	Testing Set	Validation Set	Total Images
Healthy	Grade 0	2286	639	328	3253
Doubtful	Grade 1	1046	296	153	1495
Minimal	Grade 2	1516	447	212	2175
Moderate	Grade 3	757	223	106	1086
Severe	Grade 4	173	51	27	251
Total	**All Grades**	5778	1656	826	8260

- Grade 0 (Healthy): Denotes the absence of osteoarthritis radiographic features.

- The most advanced stage, Grade 4 (severe), is characterized by severe bone deformation and joint space narrowing.

- There are much fewer photos in the severe categories (Grades 3 and 4) than in the healthy ones (Grades 0), indicating a class imbalance in the dataset. This calls for the application of Data Augmentation and Balancing techniques.

Table 3: Distribution of Annotated Images for Segmentation

Subset	Number of Images	Purpose
Training Set	876	Used to train the **Mask R-CNN** to create knee joint space polygon masks.
Validation Set	109	Used to adjust the mask heads and assess the Intersection over Union **(IoU) score**.
Total	985	Total number of photos with annotations utilized for the segmentation task.

5.2 DATA PARTITIONING AND SPLITTING STRATEGY

The dataset was carefully divided into three separate subsets: Training, Validation, and Testing, in order to guarantee a thorough assessment of the suggested deep learning models. In addition to preventing data leaks, this separation guarantees that the model's performance is assessed on unseen data, reflecting its clinical application in the real world.

5.2.1 METHODOLOGY OF SPLITTING

To preserve the distribution of disease severity across all categories, the data splitting procedure used a stratified sample technique. To enable the deep learning frameworks (TensorFlow/Keras) to automatically ingest data with accurate class labels, the **dataset** was arranged into a hierarchical directory structure as shown in Table 3.

- The model uses the largest subset, the training set (70%) to learn the feature representations of osteoarthritis in the knee. The loss computed from this data is used to modify the neural networks' weights and biases.

- **Validation Set (10%)**: During training, this subset is utilized to adjust hyperparameters (such batch size and learning rate) and keep an eye out for overfitting. In order to direct model optimization, it serves as a stand-in for test data.

- **Testing Set (20%)**: Only after model training is finished is this rigorously "held-out" subset utilized. It offers the last, objective assessment of the model's precision and capacity for generalization.

5.3 DATA ACQUISITION AND PREPROCESSING

High-quality data is the system's cornerstone. Raw medical images were converted into a deep learning-friendly format at this step.

Data Source: Mendeley Data (derived from the OAI) provides the Knee Osteoarthritis Severity Grading Dataset used by the system. 8,260 X-ray pictures with Kellgren-Lawrence (KL) grade labels are included as shown in Table 2.

Class Simplification: The original five-class system was combined into three strong categories to enhance clinical decision-making:

- **Healthy:** (Grades 0, 1, 2)
- **Moderate:** (Grade 3)
- **Severe:** (Grade 4)

5.3.1 PIPELINE FOR IMAGE ENHANCEMENT

Median Filtering: By substituting each pixel value in an image or signal with the median value of its neighbors, a median filter is a digital signal processing technique that successfully lowers noise and preserves edges. The number of neighboring pixels to examine is determined by a certain factor value. Since the median filtering "factor" is 5, a window utilized in the filtering process has dimensions of 5 pixels in width and height. This demonstrates how the median filter replaces the initial value at the neighborhood center by calculating the median value of the pixels in the 5x5 neighborhood surrounding each pixel in the image or signal. Applied to remove "salt-and-pepper" noise inherent in radiographic imaging while preserving the sharp edges of the bone structures.

For our input image $\boldsymbol{f(x,y)}$ and a filtered output image $g(x,y)$, the formula for median filter operation is defined as:

$$\boldsymbol{g(x,y) \;=\; median\,\{f(s,t) \mid (s,t) \in \; Sxy\}}$$

$\boldsymbol{g(x,y)}$ **:** The value of the pixel at coordinates (x,y) in the processed image.
$\boldsymbol{f(s,t)}$**:** The intensity values of the pixels in the original image.

$\boldsymbol{Sxy}$**:** A rectangular neighborhood (or window) centered around the pixel (x, y). In our project, this is a $5 \times 5\ times$.

Median: The statistical median function (which sorts all values in the neighborhood and picks the middle one).

In our code, we use cv2.medianBlur(img, 5). This means for every single pixel in the X-ray, the computer looks at a $5 \times 5\ square$ surrounding it (25 pixels total).

Contrast Limited Adaptive Histogram Equalization, or CLAHE, is used to improve the contrast of the **joint space** and normalize lighting, which helps to highlight the cartilage gap against the bone. CLAHE splits a picture into tiny, overlapping sub-regions and applies histogram equalization independently to each sub-region. A clip limit is a parameter that controls how much contrast enhancement can be applied to each individual pixel in an image.Two is the selected clip limit.

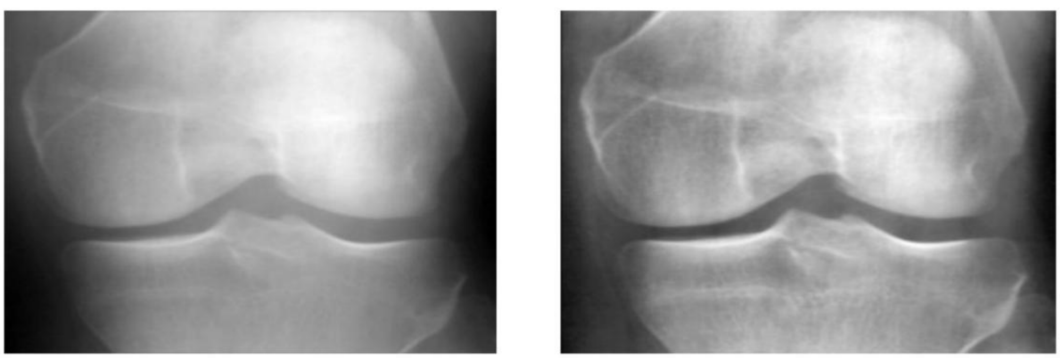

Figure 3: Before and After of Knee CALHE

An **X-ray** image with the median filter and CLAHE applied before and after is shown above. We can observe that the image is improved further after the approaches.

Data Augmentation: Data augmentation is a technique that involves applying several alterations to the current data in order to artificially expand the size of a training dataset. To address the severe class imbalance (where healthy images outnumbered severe ones), Data Augmentation was performed. Minority classes underwent random rotations, zooming, and horizontal flipping to ensure the model received a balanced number of examples for training.

Table 4: Data After Augmentation

Disease Severity	KL Grade	Original Training Images	Augmented To	Total Images (After Augmentation)
Healthy	Grade 0-2	~3000+ (Trimmed down)	Trimmed to 500	500
Moderate	Grade 3	757	Trimmed to 500	500
Severe	Grade 4	173	Augmented (Oversampled)	500
Total	All Classes	~4000+	Balanced	1500

Table 5: Model Value

HyperParameter	**Value**
Learning Rate	0.001
Number of Epochs	40
Batch Size	32
Activation Function	SoftMax
Optimization	Adam Optimizer
Dropout Rate	0.4
Loss Function	Categorical CrossEntropy

HyperParameter : Variables that must be set before a machine learning algorithm may learn them during training are known as hyperparameters. The number of hidden layers in a neural network, the speed at which the optimization algorithm learns, the strength of the regularization, and the number of iterations or epochs are some of the variables that have an impact on the learning process.

Learning Rate : This hyperparameter determines how much the model weights should change in reaction to the training error. While a higher learning rate might result in faster convergence but also overshooting the optimum weights, a lower learning rate may result in slow convergence or being trapped in local minima.

Number of Epoch : It establishes the number of times the model will be trained using the complete dataset.

Batch Size : The number of samples used in each training iteration is determined by the batch size. Larger batch sizes might reduce training times, but they might also use more memory and have poorer generalization.

Activation Function : A neural network function that helps the network identify complex patterns in the data by adding nonlinearity to each neuron's output. The activation algorithms ReLU, sigmoid, and softmax are often utilized.

Optimization : A method for identifying the best set of model parameters to minimize the training loss function. Optimization algorithms such as Adam, RMSProp, and gradient descent are commonly employed.

Dropout Rate : A regularization hyperparameter that, in order to prevent overfitting, randomly "drops out" (sets to zero) a portion of the neurons in a layer during training.

Loss Function : A function that assesses the model's performance during training on a particular task. Reducing the loss function the difference between expected and actual value is the goal.

5.4 DEVELOPMENT OF BASE CLASSIFICATION MODELS

Four separate "Base Learner" models were created during this stage in order to extract features and forecast the severity of the illness. Transfer Learning was used to make use of prior knowledge rather than starting from scratch.

Choosing a Model: Based on their unique advantages, four cutting-edge Convolutional Neural Network (CNN) architectures were chosen:

- **DenseNet121:** Selected due to its capacity for feature reuse.
- Because of its exceptional accuracy-to-parameter ratio, EfficientNetB5 was chosen.
- Because of its capacity for multi-scale processing, InceptionV3 is utilized.
- **MobileNetV2:** Because of its computational effectiveness, it is included.

Training Strategy: Each model was initialized with weights pre-trained on ImageNet. The top layers were removed and replaced with custom classification heads (Global Average Pooling ~ Dropout ~ Dense Softmax Layer). Each model was trained independently on the balanced dataset using the Adamax optimizer.

5.5 EXPLANATION OF THE MODELS & ALGORITHMS USED

5.5.1 CNN

Convolutional Neural Networks (CNNs) are a type of deep learning neural network architecture specifically designed for image and video recognition tasks. They have been actively used for a number of computer vision applications, including segmentation, object identification, and image classification. The convolutional layer, which utilizes the mathematical procedure of convolution to the input data, is the fundamental component of a CNN. The convolution operation involves sliding a small filter (also called a kernel or weight matrix) over the input data, element-wise multiplying each section of the input by the filter, and summing the results to produce a single output value. For each position of the filter across the incoming data, this step is repeated several times, producing a feature map, or set of output values. A CNN can extract increasingly complex features from the input data by stacking multiple convolutional layers on top of one another. The feature maps' spatial dimensions are reduced by the pooling layer produced by the convolutional layers, which is another crucial component of a CNN. The most prominent pooling procedure is max-pooling, which replaces the whole region of the feature map with the highest value extracted from one or more feature regions map. This makes the network more tolerant to slight fluctuations in the input data.It reduces the network's computing cost. Lastly, CNN's output layer is designed to be unique to the issue it is attempting to resolve like a sigmoid or a softmax activation function for image categorization binary classification activation function. All things considered, CNNs are among the neural network designs most frequently utilized in this field and have proven to be highly effective in various computer vision tasks.They are an effective tool for tackling a multitude of challenges due to their capacity to discover hierarchical patterns from input data and how resilient they are against changes in input variations.

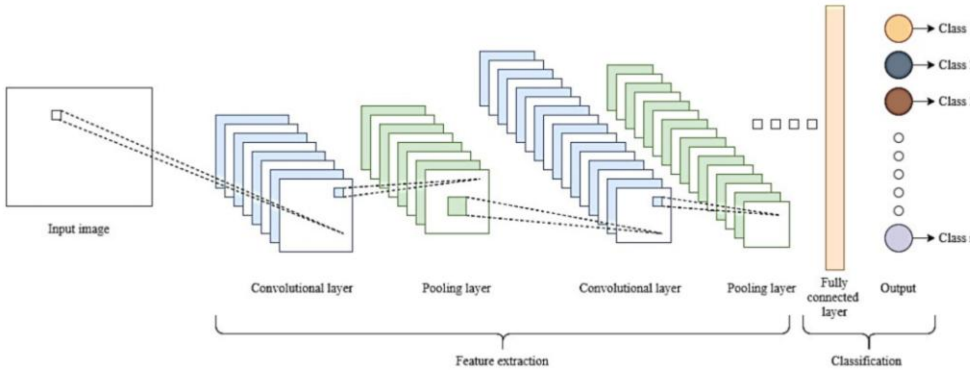

Figure 4: CNN Architecture

5.5.2 EFFICIENT NET

Google created EfficientNet, a CNN architecture used for image classification. It aims to achieve a high degree of precision with fewer parameters. The EfficientNet architecture was created to increase the network's size and resolution while maintaining accuracy and efficiency. Because of this, they can be applied to a wide range of computer vision tasks, such as image classification, object recognition, and semantic segmentation. Neural networks typically perform better with more layers. One method that can be used to do this is depth scaling. The technique known as "depth scaling" involves adding more layers to the network in order to improve accuracy, make it more powerful, and enable the identification of more features. Nevertheless, the problem of disappearing gradients arises with additional layers. A vanishing gradient is nothing, but as more layers are added, the partial derivative eventually disappears and the value of the loss function is zero. The disappearing gradient issue is typically resolved in ResNet using step connections. Step connections have now been used to solve the vanishing gradient problem, although this involves a lot of computation and other work due to the large number of layers. Therefore, in addition to depth, we additionally scale the width and resolution in EfficientNet. Resolution scaling is the process of increasing the number of pixels; depth scaling is the process of increasing the number of channels; and breadth scaling is the process of increasing the number of feature maps. More features can be extracted from the image by increasing the number of feature maps, and more features can be learned and the algorithm performs better by scaling the resolution. In order to prevent the vanishing gradient issue, we uniformly scale the depth, width, and resolution. The scaling factor, or the amount of scaling needed, is now necessary. Nowadays, efficientNet is the only implementation of a technique known as compound scaling that accomplishes this. The process of compound scaling is as follows:

$$f = d.w^s.r^s$$

The variables f, d, w, and r stand for how much we can scale, depth, width, and resolution scaling factors, respectively. The value of s, which is the result of grid search, is 1.

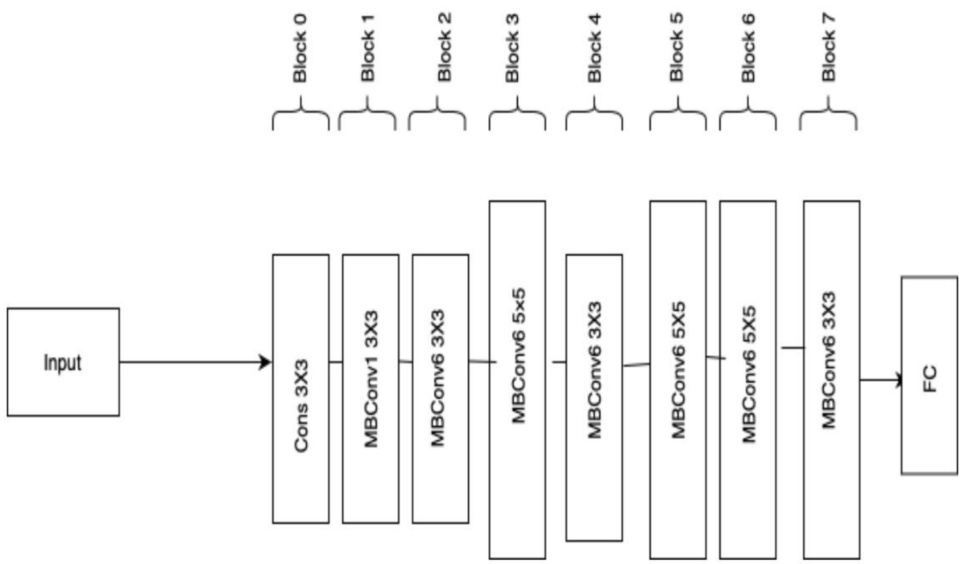

Figure 5: Efficient NET Architecture

5.5.3 INCEPTION V3

In order to improve the effectiveness and precision of picture classification tasks, Google researchers created the Inception-v3 convolutional neural network (CNN) architecture, which combines convolutional layers with varying filter sizes to gather features at different scales. A variety of convolutional, pooling, and other layer types comprise each module of the Inception v3 architecture. These modules are arranged in a hierarchy, where higher-level modules focus on capturing more complicated features like object components and entire objects, while lower-level modules concentrate on capturing low-level features like edges and textures. This network's modules consist of a max-pooling layer that aids in lowering the dimensionality of the output and parallel convolutional layers with different filter sizes. The output of each module is then aggregated and sent to the subsequent module. Additionally, batch normalization and dropout regularization are used in Inception-v3. All things considered, it has been shown that Inception v3 excels in a number of picture recognition tasks, such as visual question answering, object detection, and image classification. Furthermore, this has been used to extract characteristics that have shown great efficacy for transfer learning tasks.

InceptionV3's architecture

• **Input layer:** The layer that gets the raw image data and sends it to the layer after it.

• **Factorized convolutions:** InceptionV3 replaces traditional convolutions with factorized convolutions, which are more computationally efficient. Factored convolutions split the standard convolution into two smaller convolutions, one along the width of the input and one along its height. This speeds up network training and inference by lowering the network's parameters.

• **Inception Modules:** Inception modules are the basis of the InceptionV3 architecture. They consist of several parallel convolutional layers of different sizes, pooling layers, and 1x1 convolutional layers. The purpose of these modules is to gather features at different scales and combine them into a single layer.

• **Improved pooling:** InceptionV3 replaces the more traditional "max pooling" technique with "global average pooling." Global average pooling reduces overfitting and is more resilient to small changes in input. Additionally, the number of network parameters is decreased as shown in Figure 6.

• **Auxiliary classifiers:** InceptionV3 features two auxiliary classifiers connected to the intermediate layers of the network. These auxiliary classifiers provide the network with additional training signals during training in an effort to improve overall accuracy and reduce overfitting.

• **Batch normalization:** InceptionV3's batch normalization layers help to lower the network's sensitivity to the initial weight values and stabilize the learning process.

• **Stem network:** InceptionV3's architecture incorporates a "stem network" designed to extract fundamental features and preprocess input images. This helps to improve network performance and reduce the number of parameters.

• **Reduction layers:** InceptionV3 features "reduction layers" that are used to reduce the spatial dimensions of the feature maps while increasing the number of channels. This helps to improve performance and reduce the computational complexity of the network As shown in Figure 6.

• **Fully connected layers:** At the network's end, the fully connected layers do the last categorization of the input image. InceptionV3 has two fully connected layers, one with 2048 units and the other with 1000 units.

• **Softmax layer:** Using the output of the last fully linked layer, the softmax layer generates a probability distribution across the classes.

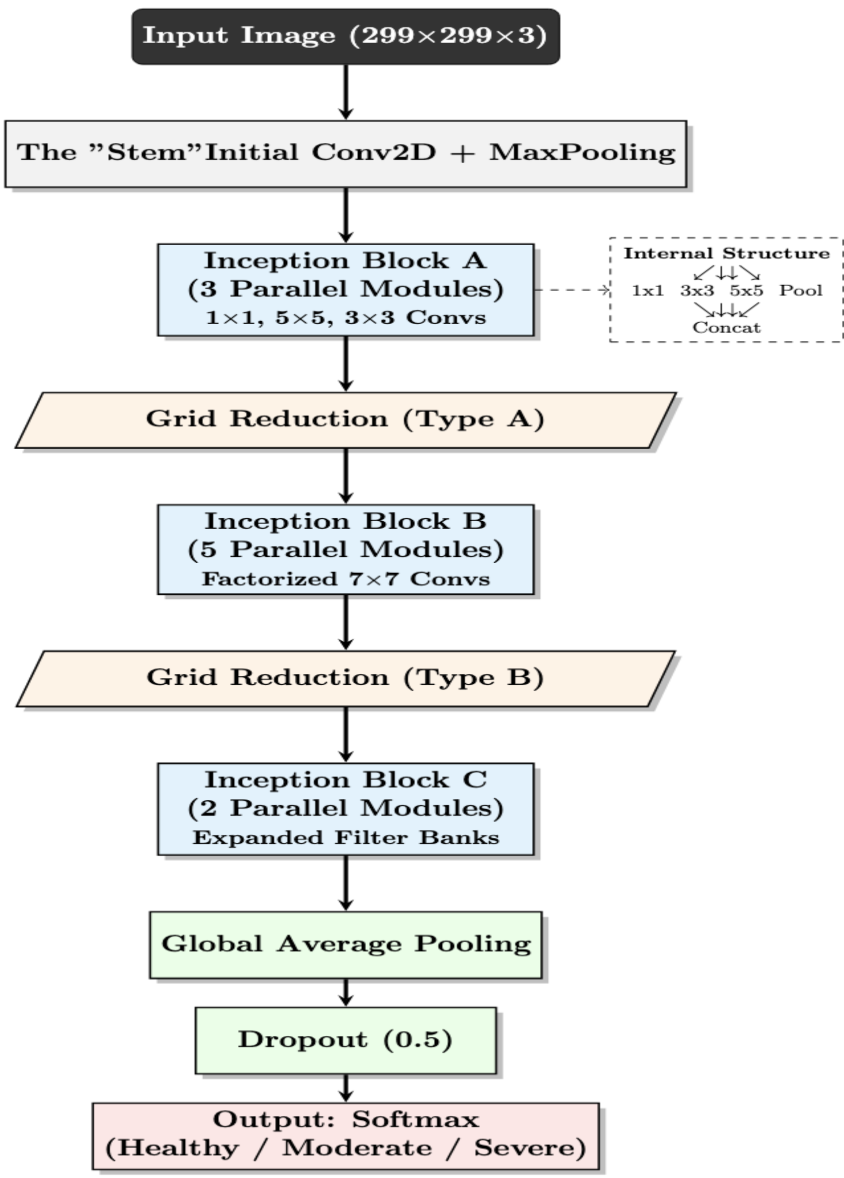

Figure 6: Architecture of Inception V3

5.5.4 DENSE NET

We integrate DenseNet121, a design that improves feature propagation through densely connected layers, into our ensemble framework. The capacity of DenseNet to reuse features across layers is very useful for medical imaging tasks where subtle patterns are essential, as demonstrated by Muhammad et al.[4] in their ensemble model for knee osteoarthritis assessment. The efficacy of including such sophisticated CNN architectures in ensemble models to enhance knee osteoarthritis diagnostic accuracy is further validated by Bhateja et al. [5]. We hope to create a more reliable and accurate severity classification system by utilizing DenseNet121 in our ensemble to capture complex joint degradation characteristics while preserving computational efficiency.

DenseNet is composed of several dense blocks, each of which is composed of multiple dense layers. Each dense layer consists of a convolutional layer sequence with the same number of filters, a batch normalization layer, and an activation function called a Rectified Linear Unit (ReLU). The outputs from all of the dense block's earlier layers are pooled and sent into the current layer as input. The output of each dense block is subsequently sent via a transition layer, which performs a number of operations to minimize the quantity and spatial dimensions of the feature maps. The transition layer consists of a 2x2 average pooling layer with stride 2, a 1x1 convolutional layer with a reduction factor (often set to 0.5), and a batch normalization layer as shown in Figure 7.

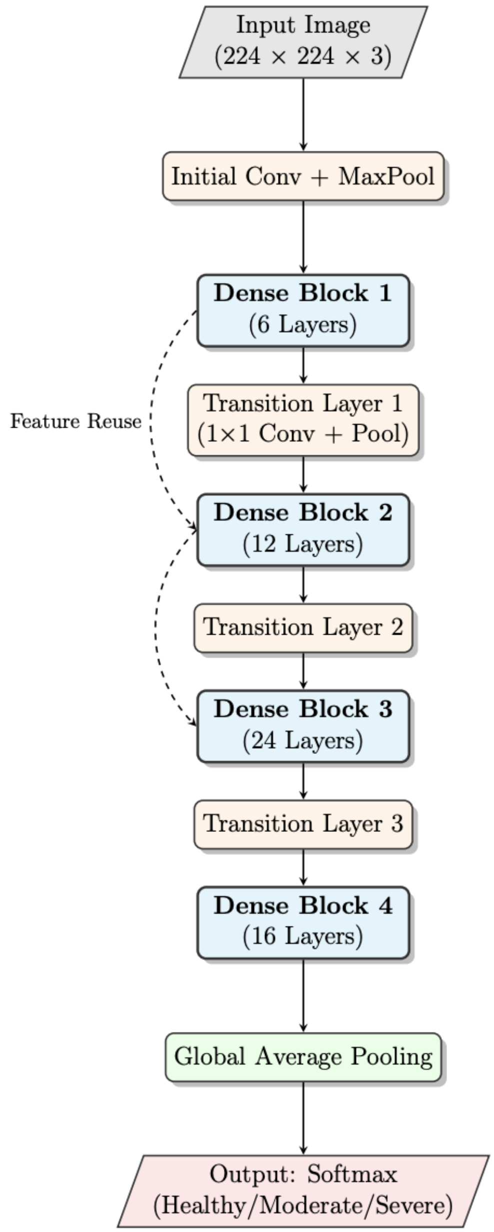

Figure 7: Dense Net Architecture

5.5.5 MOBILE NET

Google researchers developed the CNN architecture known as MobileNet, which makes use of depthwise separable convolutions to reduce the network's processing overhead without significantly decreasing accuracy. The idea underlying depthwise The purpose of separable convolutions is to replace two separate processes, a depthwise For the standard convolution technique, convolution and a pointwise convolution which combines a filter tensor with the input tensor. The convolution in depth generates a collection of output channels by applying a spatial convolution to each input channel with a small filter kernel, often 3x3 or 5x5 as shown in Figure 8. The depth-wiseAfter then, the result of the convolution undergoes a 1x1 convolution in the pointwise convolution, which uses a linear filter to combine the output channels. This network may learn complex information while using fewer computations characteristics due to the depthwise separable convolutions that are used. This results in MobileNet being a great choice for devices with constrained resources, such as embedded systems and smartphones.All things considered, MobileNet is a compact and lightweight CNN design that works well.making it ideal for embedded and mobile devices. Additionally, it can be applied to various applications with constrained processing power.

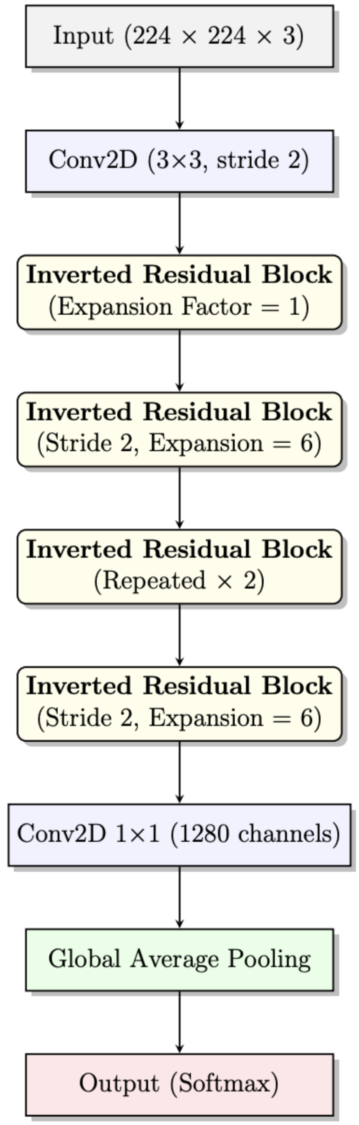

Figure 8: MobileNet Architecture

5.5.6 MASK R-CNN

A deep learning model for instance segmentation and object detection is called Mask R-CNN.The Mask R-CNN model builds on the Faster R-CNN model by adding a branch to predict segmentation masks for every detected object. Consequently, in addition to identifying the objects in the image, the model can also segment objects in a picture, that is, determine which pixels belong to an item and which do not. In scholarly publications, Mask R-CNN can be utilized as a foundational or state-of-the-art model for a variety of computer vision issues, such as semantic segmentation, instance segmentation, and object identification. Researchers can use pre-trained Mask R-CNN models or train their own models on their own datasets to evaluate the efficacy of their proposed methods or models. Additionally, the pre-trained Mask R-CNN model can be used by researchers to extract characteristics from images that can be used in downstream models or other tasks. This procedure is referred to as transfer learning or feature extraction. On several object detection and segmentation benchmarks, such as COCO and Cityscapes, Mask R-CNN has demonstrated state-of-the-art performance. Numerous fields have made use of it, including robots, medical imaging, and autonomous driving.

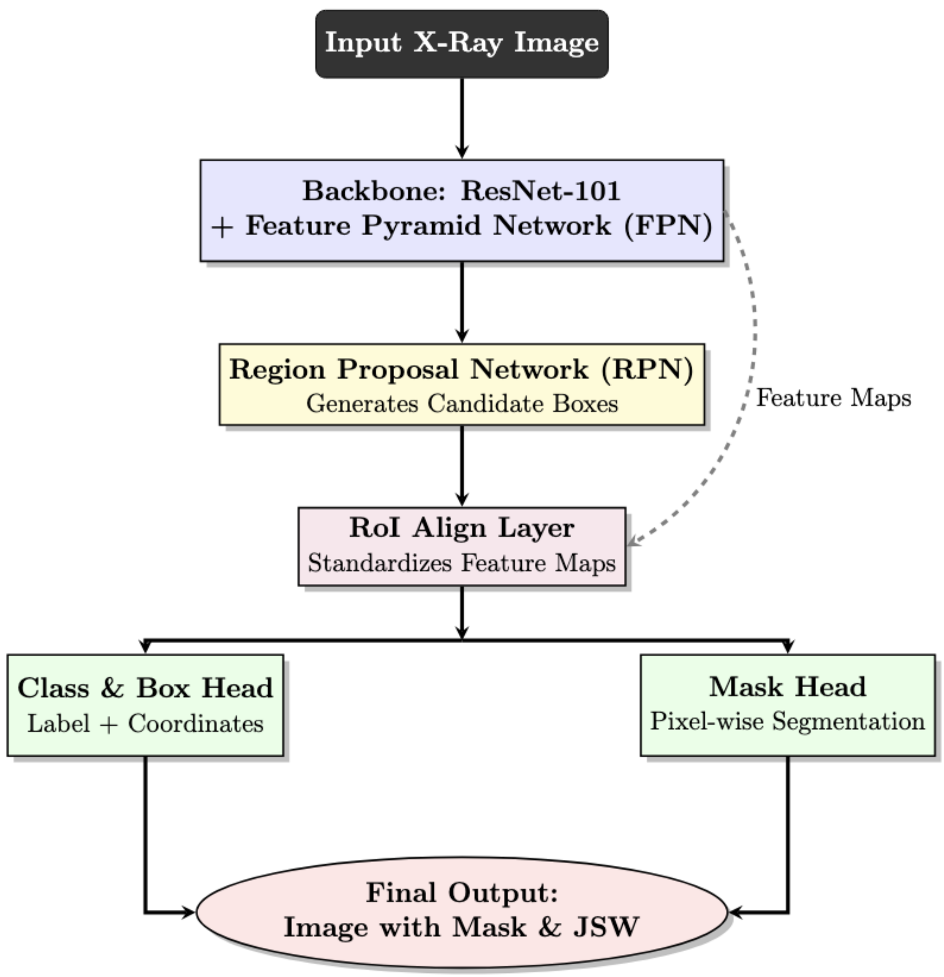

Figure 9: Mask R-CNN Architecture

5.6 ENSEMBLE MODEL INTEGRATION

This phase concentrated on merging the base learners into a single, reliable predictor in order to reduce the bias and variation of separate models.

- **The strategy of an ensemble:** The method of Averaging Ensemble (Hard Voting) was used.
- **Procedure:** All four trained models are simultaneously run through the input X-ray. A probability distribution for each of the three classes—Healthy, Moderate, and Severe is produced by each model.
- **Aggregation:** The system determines these probability' arithmetic mean. The final forecast is chosen from the class having the highest average probability. When compared to a single model, this "wisdom of the crowd" method dramatically decreased misclassification errors.

5.7 SEMANTIC SEGMENTATION WITH MASK R-CNN

This stage concentrated on localization determining the precise location of the joint space while classification offers a grade.

- **Model Architecture:** Mask R-CNN (Region-based Convolutional Neural Network) was deployed for instance segmentation.
- **Annotation:** Polygons indicating the limits of the tibia, femur, and joint space between them were manually added to a portion of the data.
- **Training:** The model was trained to produce a binary pixel mask specifically for the knee joint gap (Region of Interest) while ignoring the background. This stage successfully "teaches" the computer to differentiate between bone and cartilage/space.

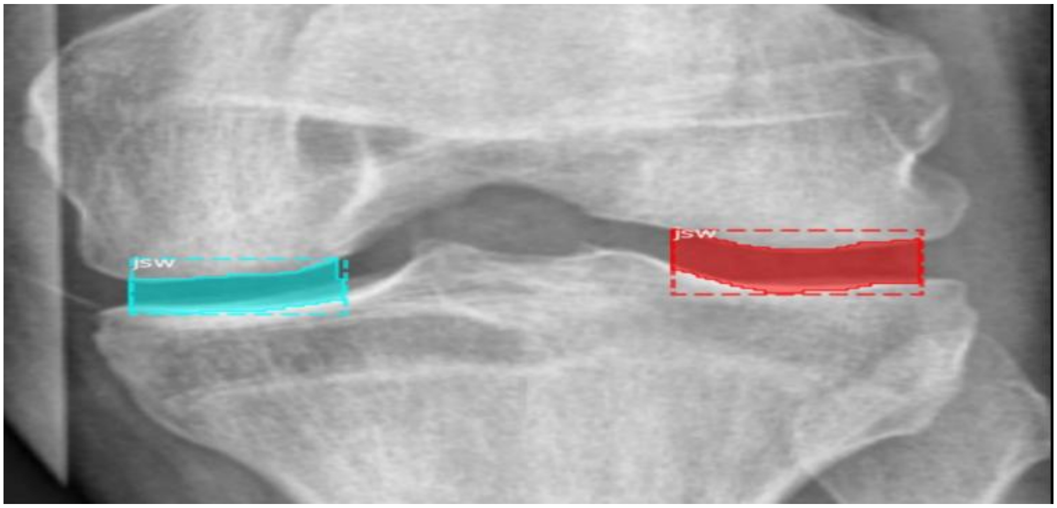

Figure 10: Region Captured by Mask R-CNN

5.8 QUANTITATIVE JOINT SPACE WIDTH (JSW) MEASUREMENT

The visual mask created in Phase 4 is transformed into a clinical numerical value in this phase.

- **Centroid Detection:** To determine the geometric center (centroid) of the femoral and tibial borders, the system examines the binary mask.
- **Distance Calculation:** The vertical distance in pixels between these centroids is determined using a Euclidean distance technique.
- **Metric Conversion:** The pixel distance is converted to millimeters (mm) using a calibration factor (pixels per millimeter). The ultimate Joint Space Width (JSW) is produced in this way, giving the physician an objective figure (such as "2.45 mm") to monitor the course of the illness.

5.9 MEDICAL VALIDATION OF THRESHOLDS

We worked with **Dr. Mudit Sharma(M.S. Orthopaedics)** who is currently working as Orthopedic Surgeon in New Delhi to determine the key thresholds for Joint Space Width (JSW) in order to guarantee the diagnostic value of our automated assessments. The following classification criteria were established for our system based on his clinical experience and accepted orthopedic protocols:

Healthy: JSW $> 4mm$

Early/Moderate OA: JSW between $2 - 4mm$

Severe OA: JSW$< 2mm$

This medical consultation guarantees that our segmentation module's output complies with hospital-standard radiological assessment procedures.

5.10 PROPOSED SYSTEM DESIGN ARCHITECTURE

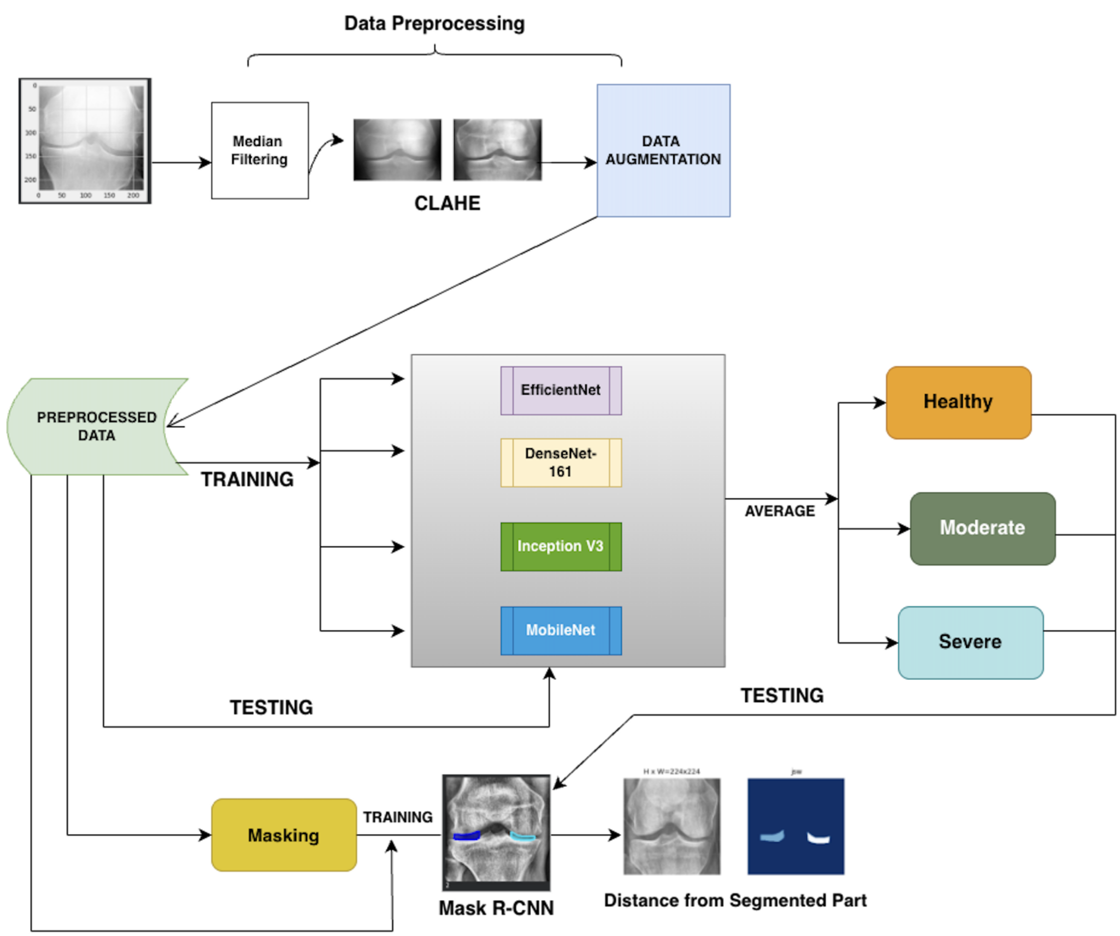

Figure 11: Overall Architecture of Knee OsteoArthritis

CHAPTER 6: TESTING AND RESULTS

6.1 LIBRARIES AND TECHNOLOGIES USED

Tensorflow : An open-source framework for building and implementing machine learning models is called TensorFlow. Developed by Google, it is extensively used in both business and education. TensorFlow allows users to create and train neural networks and provides resources for implementing models in practical contexts.

Keras: Developed in Python, Keras is a high-level neural network API that may be utilized with TensorFlow, CNTK, or Theano. Keras makes building and experimenting with neural networks easy with its intuitive interface for building models and training neural networks.

NumPy: The Python package Array manipulation is done with NumPy. It provides tools for working with arrays in addition, multiplication, and dot product operations. NumPy is widely used in machine learning, scientific computing, and data analysis.

Pandas: A Python package for data manipulation and analysis. The tools offered for working with tabular data, such as data frames, allow users to do tasks like grouping, filtering, and merging.

Matplotlib: A visualization library for Python. It provides tools for creating a wide variety of graphs, such as bar charts, scatter plots, line plots, and more. Matplotlib is widely used in data analysis and scientific computing.

OpenCV : A computer vision library for processing images and videos is called OpenCV, or Open Source Computer Vision Library. It provides tools for tasks including object detection, image segmentation, and face recognition.

SeaBorn: To create statistical visualizations, a Python program named Seaborn is utilized. It provides tools for creating a wide range of graph types, such as heatmaps and histograms. Seaborn is widely used in data analysis and machine learning.

6.2 DETAILED IMPLEMENTATION OF EACH MODULE

6.2.1 DATA PREPROCESSING

Preprocessing, or the preparation and transformation of raw data into a more usable form, is a crucial step in data analysis and machine learning. To make the data more suitable for analysis and modeling, data transformation, normalization, and cleansing are frequently used.

Making sure the data is accurate, consistent, and suitable for the specific analysis or machine learning task at hand is the primary goal of data preparation. In addition to addressing issues like missing data, outliers, inconsistencies, and formatting errors, this frequently entails determining and changing the most important components or variables in the data.

Overall, data preparation is a crucial step in the data analysis process since it has a significant impact on the data's quality and accuracy. Our preprocessing methods include data augmentation, CLAHE, and median filtering.

6.2.2 TRAINING AND ENSEMBLE

Figure 12: Class Prediction by Ensemble Model

All of the models are trained using the preprocessed photos. Each picture is 224 by 224 pixels. Each model's entire train dataset is separated into batches of 32. After 20 epochs, the learning rate, which began the training process at 0.001, is progressively reduced to **0.0005**. The Adam optimizer is used for training. The separate models are then combined to create an ensemble model. Every model generates a probability distribution for image classification. By averaging the probability distributions and selecting the class with the highest probability, the ensemble model aims to aggregate the findings from every model. A different dataset is used to train and assess the ensemble model as shown in Figure 12.

6.2.3 FINDING DISTANCE

CREATING MASK : Using the polygon tool, we manually design a mask for each X-ray image so that the region of interest may be found.

USING MASK FOR TRAINING THE DATA :

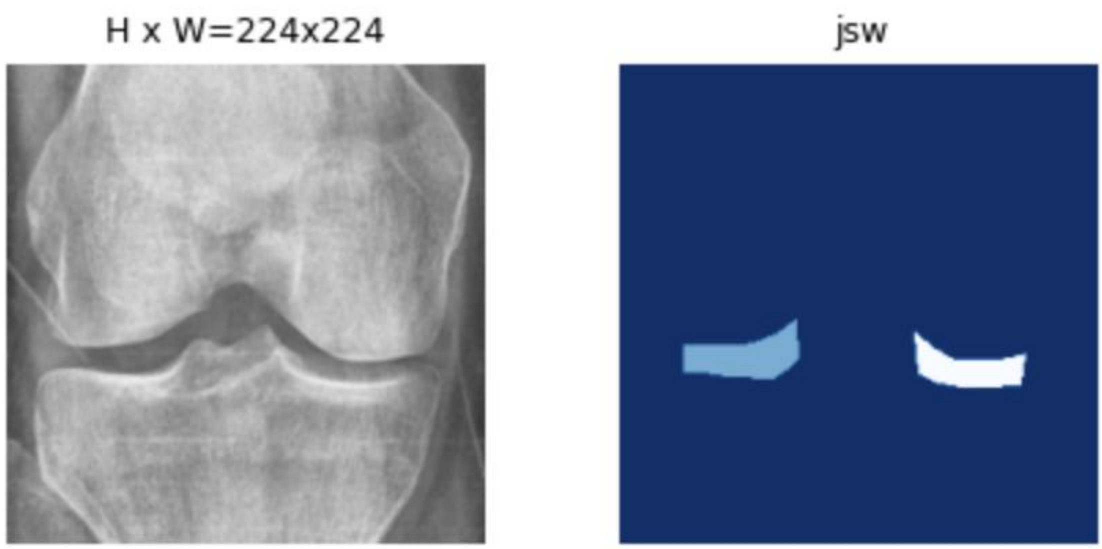

Figure 13: Image Sample of Mask R-CNN

For segmentation, we employ the Mark R-CNN architecture, which is trained on X-ray images using a mask.

FINDING JOINT SPACE WIDTH :

Prior to determining the points above and below the centroid, we first determine the centroid of the masked region.The distance is determined once the top and bottom points have been identified.After locating the centroid's left and right points, the procedure is repeated.The minimum distance is now selected.

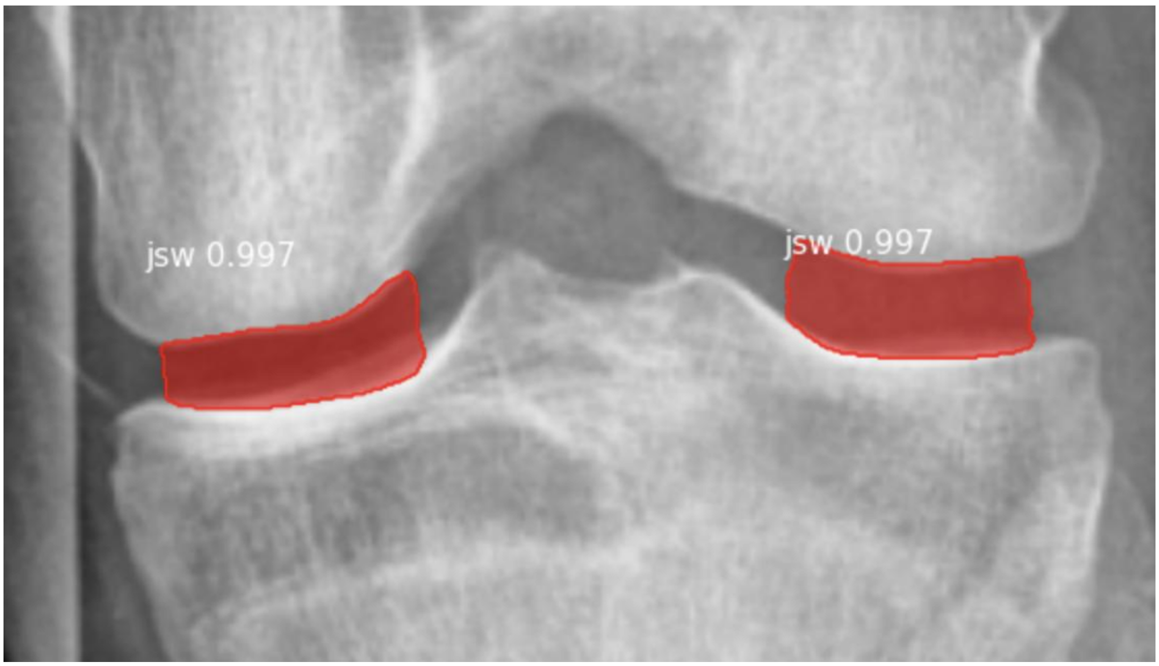

Figure 14: JSW Area Measurement

6.3 PERFORMANCE EVALUATION

Metric Employed : Precision, recall, F1 score, support, and accuracy for classification are the performance metrics used to assess the model's performance.The segmentation performance metric is intersection over union.

Precision : A measure that assesses the proportion of correctly detected positive samples among all positive samples.

$$\boldsymbol{Precision = \frac{TP}{TP + FP}}$$

Recall : A statistic that calculates the proportion of samples that are actually positive among all samples.

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}}$$

F1 Score : A test that combines recall and precision into a single score in order to balance the two. The following formula is used to calculate the harmonic mean of precision and recall:

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

Support : The quantity of samples in every class. This is used to calculate the weighted average of the performance metrics for multi-class classification issues.

$$\boldsymbol{Support} = \frac{\boldsymbol{Frequency}(\boldsymbol{x},\boldsymbol{y})}{\boldsymbol{N}}$$

Accuracy : A statistic that determines the proportion of all samples that were successfully identified.

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

Table 6: Model Performance

Model Architecture	Accuracy	Precision	Recall	F1-Score
DenseNet121	94.87%	95.03%	94.87%	94.87%
EfficientNetB5	94.14%	94.49%	94.14%	94.27%
InceptionV3	94.08%	93.87%	94.08%	93.94%
MobileNetV2	93.66%	93.85%	93.66%	93.73%
Ensemble Model	99.03%	99.02%	99.03%	99.02%

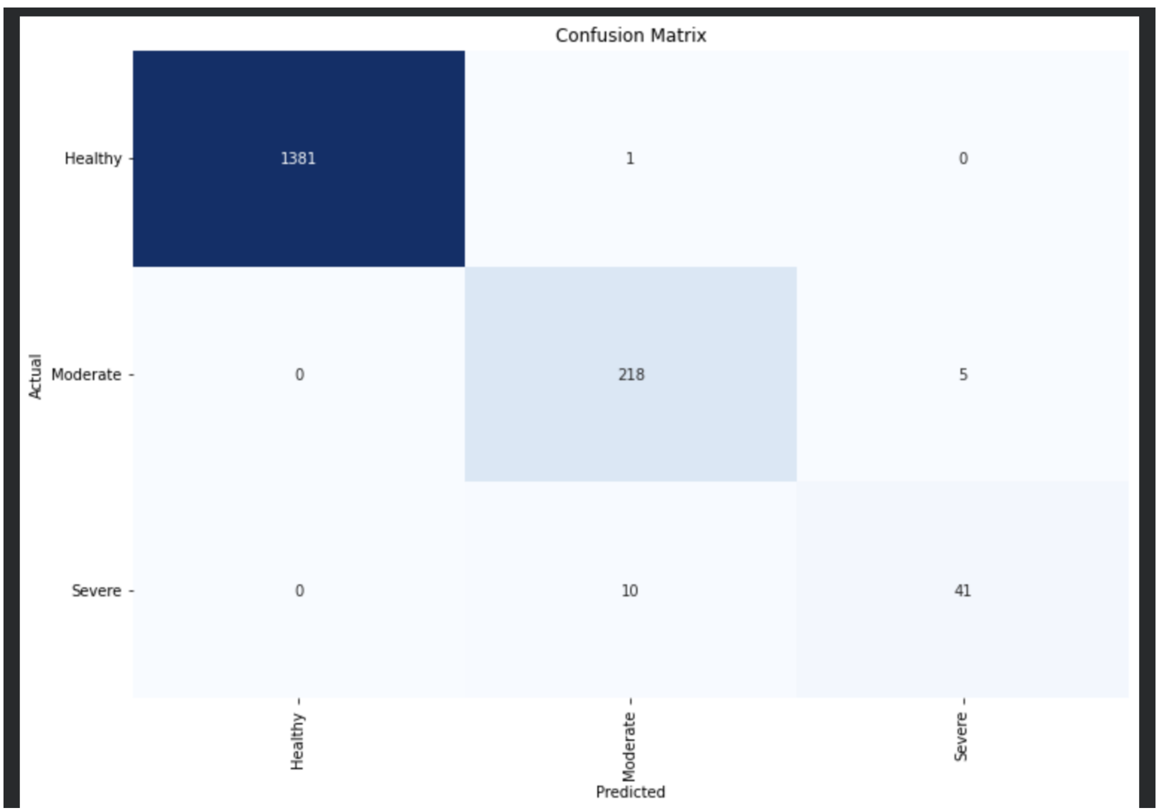

Figure 15: Confusion Matrix Of Ensemble Model

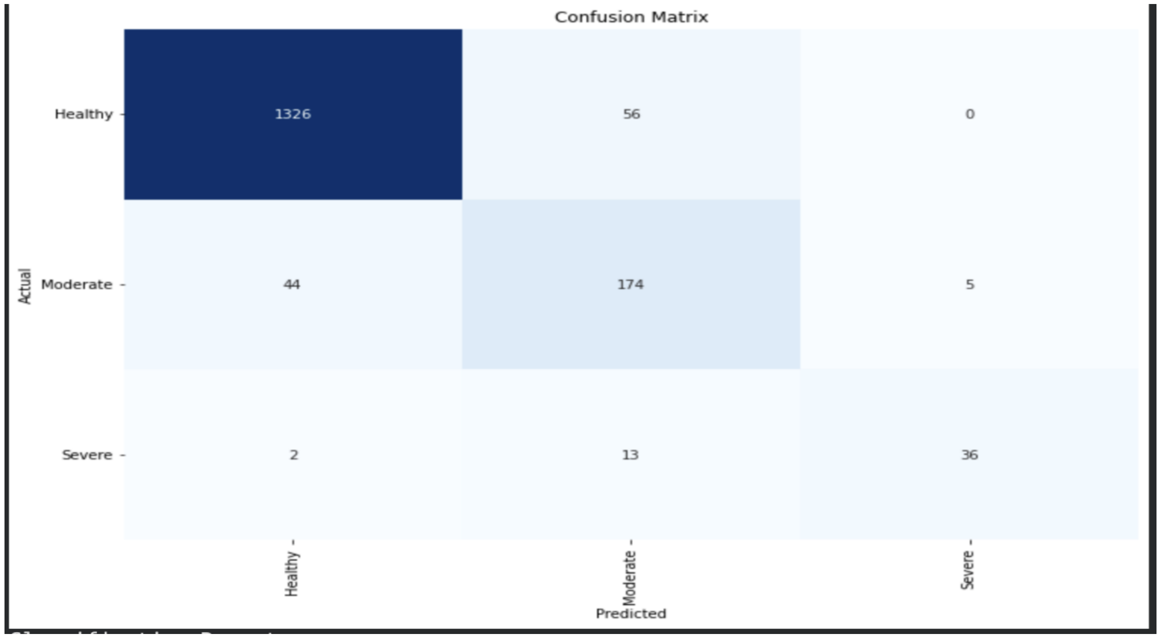

Figure 16: Confusion Matrix Of EfficientNetB5

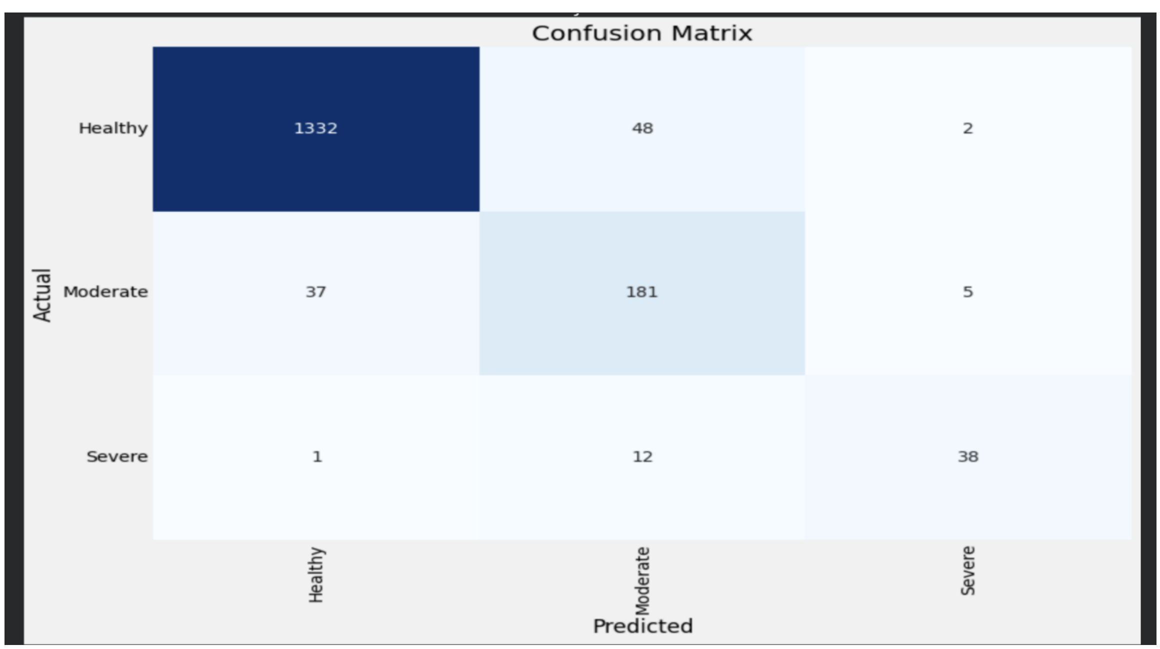

Figure 17: Confusion Matrix Of MobileNetV2

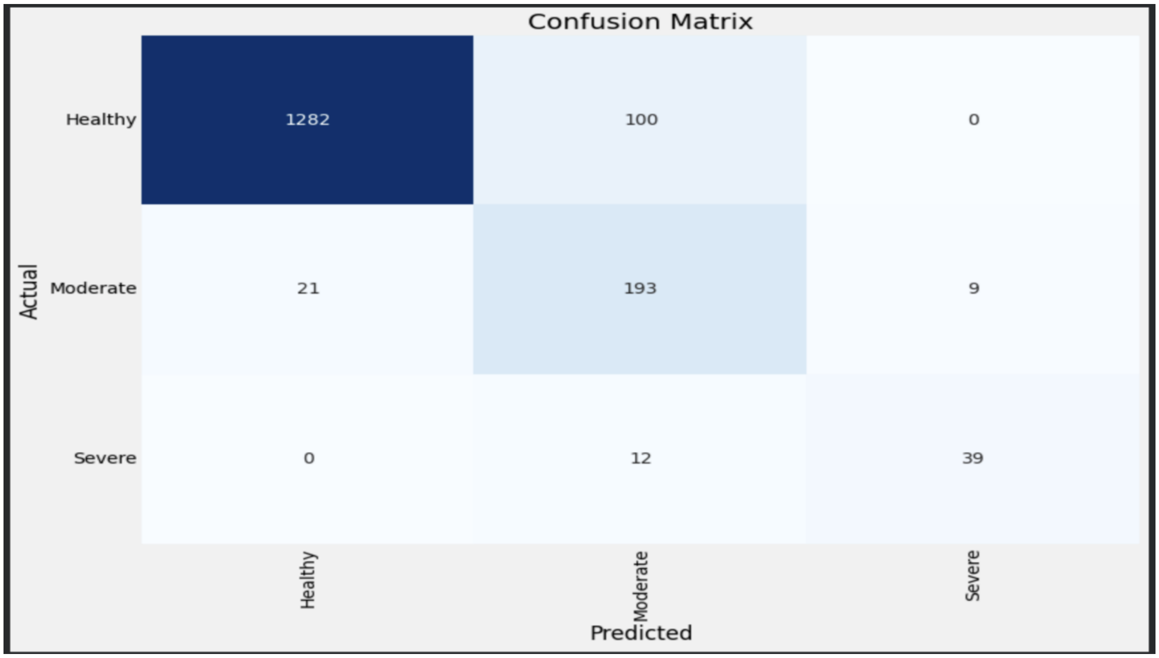

Figure 18: Confusion Matrix Of InceptionV3

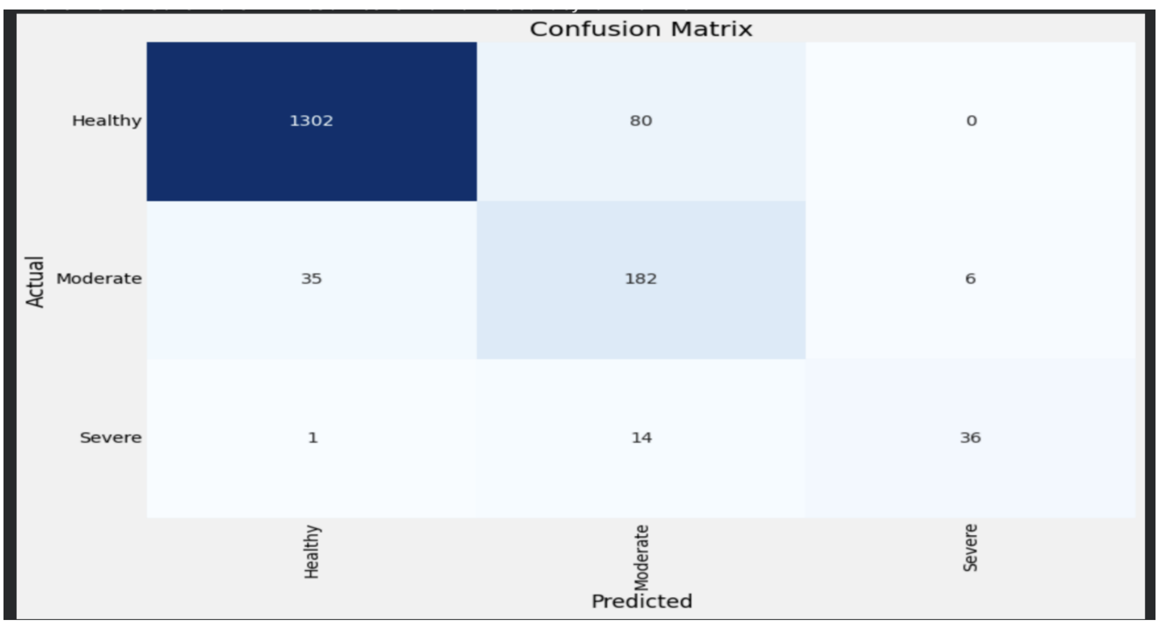

Figure 19: Confusion Matrix Of DenseNet121

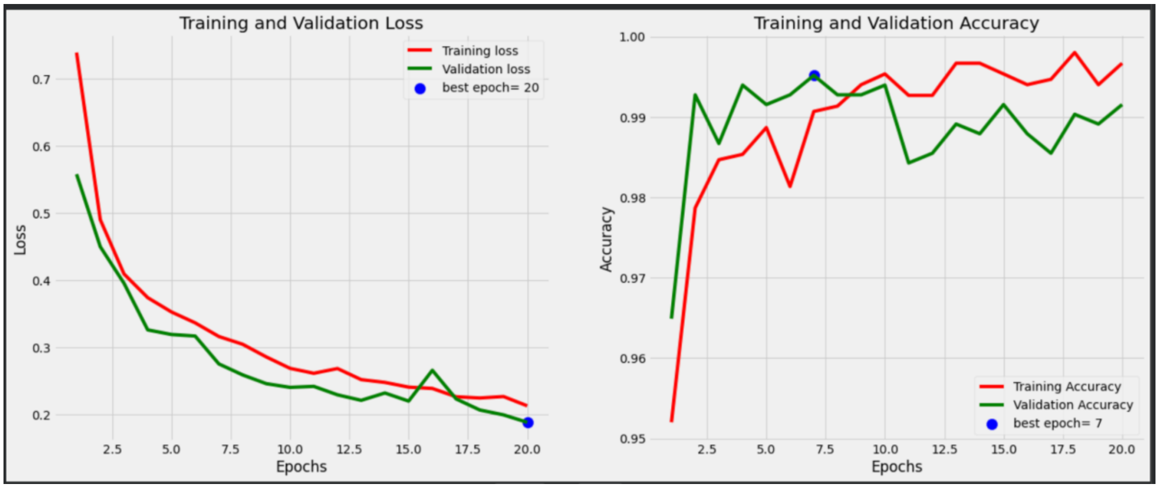

Figure 20: Training and Validation Loss

6.4 INTERSECTION OVER UNION

An object detection or segmentation algorithm's accuracy is assessed using the Intersection over Union (IoU), commonly known as the Jaccard Index. The overlap between the ground truth and the predicted bounding box or segmentation mask is calculated as dividing the regions of the intersection and union of the two. We obtained an IOU score of 0.7321 using IOU as the segmentation metrics.

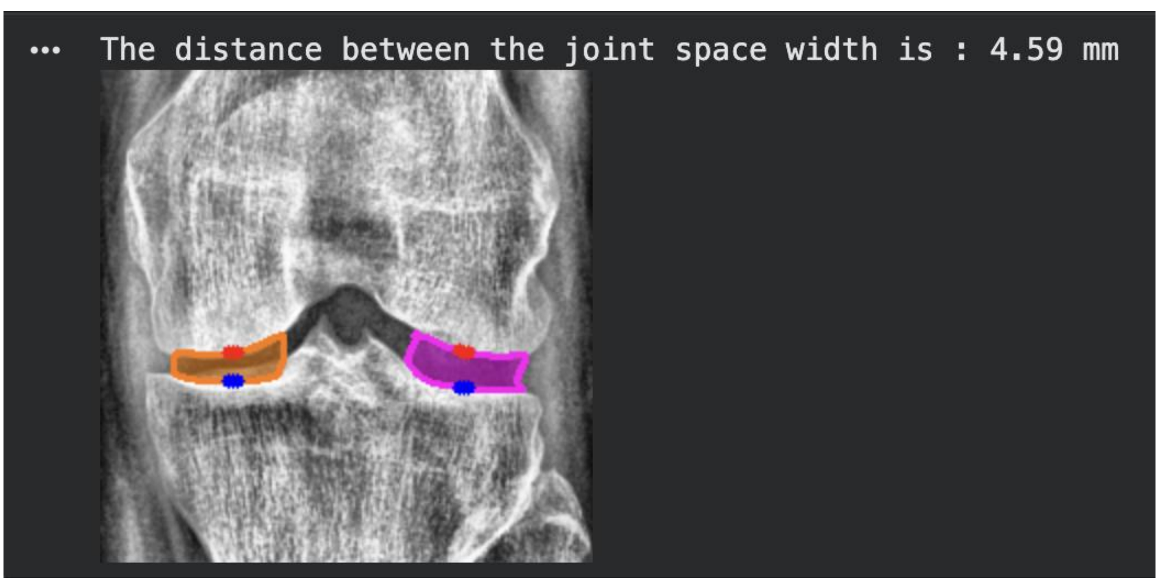

Figure 21: Final Output

CHAPTER 7: LIMITATIONS

The suggested automated system for evaluating Knee Osteoarthritis shows great accuracy and usefulness in clinical settings, but it has some problems because of the way the study was set up, the characteristics of the dataset, and the way the computer works. Recognizing these limitations is crucial for comprehending the system's present scope and delineating future research trajectories.

7.1 COMPUTATIONAL COMPLEXITY AND RESOURCE INTENSITY

The main trade-off for getting high classification accuracy is that it costs more to do so.

Inference Latency: The Ensemble Model needs to run four different deep neural networks (DenseNet121, EfficientNetB5, InceptionV3, and MobileNetV2) for each prediction. This parallel processing makes the inference time a lot longer than using just one model. **Hardware Dependencies:** Adding Mask R-CNN for segmentation puts a lot of strain on the computer. Using this dual-stream system (Ensemble + Segmentation) in a real-time clinical setting, like on a regular hospital PC or mobile device, would probably need high-end GPUs or cloud-based processing. This may not be possible in environments with limited resources.

7.2 LIMITATIONS OF 2D RADIOGRAPHY

The system only uses plain X-ray images, which naturally show a 3D anatomical structure on a 2D plane.

Loss of Depth Information: Regular X-rays can't show how the cartilage has lost volume. From a coronal view (front-facing), a joint might look wide enough, but it could have a lot of cartilage loss in the back, which an MRI or CT scan would show but our model would not.

Sensitivity to Patient Positioning: The Joint Space Width (JSW) measurement is very sensitive to the angle of the X-ray. If a patient's knee is bent or turned slightly, the 2D projection of the joint space may look like it is too narrow or too wide. This could cause the Mask R-CNN to make wrong JSW calculations.

7.3 DATASET CONSTRAINTS

The model was trained on a large dataset from the Osteoarthritis Initiative (OAI), but there are still some problems with the data.

Class Imbalance: Even though we used data augmentation methods like rotation and flipping, the original dataset was still very biased toward "Healthy" and "Early-stage" cases, with a lot fewer "Severe" (Grade 4) cases. This imbalance is a common problem with medical datasets, and it could make the model less reliable when it comes across very rare or complicated deformities that happen in severe cases.
Demographic Bias: If the training data doesn't include a wide range of ages, genders, or ethnicities, the model's predictions might not work for people who aren't in the OAI cohort.

7.4 LACK OF CLINICAL CONTEXT (UNIMODAL ANALYSIS)

The current system functions as a unimodal framework, examining solely the visual image data.
Missing Risk Factors: Osteoarthritis is a disease with many causes, including age, Body Mass Index (BMI), past injuries, and how much pain the patient says they are in. Our present model fails to incorporate these essential clinical variables.
Symptom Mismatch: The literature Alexos et al.[15] shows that there is often a difference between how bad the X-ray shows the problem is and how bad the patient feels. A patient may exhibit "Severe" structural damage while experiencing minimal pain, or conversely. Our system can't tell what the patient is really feeling or what kind of pain management they need because it only looks at the picture.

7.5 DEPENDENCE ON PREPROCESSING

The system's high accuracy is closely related to the specific preprocessing pipeline (Median Filtering and CLAHE). The model's performance is not assured on unprocessed, noisy, or low-contrast images acquired from various X-ray machines that have not received uniform enhancement. This makes it so that the system only works properly if the "input quality" is carefully controlled.

CHAPTER 8 : CONCLUSIONS AND FUTURE WORK

8.1 CONCLUSION

This study effectively sought to reconcile the significant disparity between subjective visual diagnosis and objective quantitative evaluation in orthopedics. We tackled the two biggest problems with managing Knee Osteoarthritis (KOA) by making a Dual-Stream Computer-Aided Diagnosis (CAD) System. These problems are the inconsistency of manual grading and the lack of accurate morphological measurements.

We used an Ensemble Deep Learning Model (which combined DenseNet121, EfficientNetB5, InceptionV3, and MobileNetV2) to classify severity. It worked very well, getting 99.03% accuracy on the test set. This is a lot better than single-model architectures, which supports the idea that ensemble voting can reduce bias in individual models. At the same time, the Mask R-CNN Segmentation Module successfully automated the extraction of the Region of Interest (ROI) and calculated the Joint Space Width (JSW) with high geometric accuracy (IoU 0.732).

This system is important because it doesn't just copy a radiologist's grade; it adds useful information to it. The tool gives a clear "second opinion" by giving both a severity class (like "Moderate") and a specific millimeter measurement (like "2.1 mm"). This helps doctors make decisions based on evidence for early intervention.

8.2 FUTURE SCOPE

The current system sets a strong standard for knee analysis, but there is a lot of room for growth. Our future plans include turning this prototype into a universal, multi-modal diagnostic platform.

8.2.1 GENERALIZATION ACROSS MULTIPLE JOINTS (HIP AND HAND OA)

The Kellgren-Lawrence (KL) grading scale is the best way to measure Hip and Hand OA.

- **Proposed Extension:** We intend to reeducate our Ensemble and Mask R-CNN architectures using datasets for Hip (identifying joint space narrowing in the acetabular-femoral joint) and Hand (evaluating interphalangeal joints).
- **Goal:** To make a single "Multi-Joint Diagnostic Engine" that can automatically find the body part (Knee, Hip, or Hand) and use the right grading logic.

8.2.2 INTEGRATION OF CLINICAL DATA IN REAL TIME

OAI and other academic datasets are not dynamic. We want to close the gap between the lab and the hospital to make sure that clinical work is strong.

- **Action:** We plan to work with partner hospitals to get a steady stream of real-time patient data. This will let us see how well the model holds up against the "noise" of real-world clinical settings, like different X-ray machines and how patients are positioned.
- **Advantage:** The model will be able to adapt to local demographics and rare pathological variations because it will keep learning from new hospital cases.

8.2.3 MULTIMODAL IMAGING (X-RAY AND MRI)

X-rays are great for seeing bone structure, but they don't show damage to soft tissue.

- **Proposed Extension:** We want to add Magnetic Resonance Imaging (MRI) datasets to our pipeline. MRI shows cartilage thickness, meniscal tears, and bone marrow lesions much better than other tests.
- **Goal:** To create a "Hybrid Fusion Model" that combines structural JSW data from X-rays with volumetric cartilage data from MRI to give a full 3D picture of joint health that 2D X-rays can't give on their own.

8.2.4 MAKING A UNIVERSAL DIAGNOSTIC APP

The main goal is to make things easy to get to. Our goal is to turn this complicated backend into a frontend app that is easy to use.

- **Proposed Product:** An orthopedic clinic app that works on both the web and mobile devices.

- **Functionality:** The app will have a **"Intelligent Upload Interface"** that

1. **Automatically detects modality:** It can tell if the input is an X-ray or MRI.
2. **Auto-Detects Region:** Tells us if the picture is of a knee, hip, or hand.
3. **Instant Analysis:** Runs the right segmentation model to give the JSW and Severity Grade in seconds, without the user having to choose parameters by hand.

We want to turn this project from a specialized research tool into a complete, scalable platform that can standardize osteoarthritis care across different joints and imaging modalities by following these steps.

CHAPTER 9 : REFERENCES

[1.] Sohail, M., Azad, M. M., & Kim, H. S. (2025). Knee osteoarthritis severity detection using deep inception transfer learning. Computers in Biology and Medicine, 186. https://doi.org/10.1016/j.compbiomed.2024.109641

[2.] Panwar, P., Chaurasia, S., Gangrade, J., & Bilandi, A. (2025). Early diagnosis of knee osteoarthritis severity using vision transformer. BMC Musculoskeletal Disorders, 26(1). https://doi.org/10.1186/s12891-025-09137-2

[3.] Yoshikawa, R., Segal, N. A., Tolstykh, I., Ho, M., Anderson, D. D., Lynch, J. A., Duryea, J., & Nevitt, M. C. (n.d.). Usefulness of 3D Joint Space Width on Weight-Bearing CT in Comparison with 2D Joint Space Width on Radiographs for Predicting 24-Month Worsening of Knee Osteoarthritis Pain and Function in the MOST Study HHS Public Access. https://www.elsevier.com/about/policies-and-standards/sharing

[4.] Amjad, J., Sajid, M. Z., Amjad, A., Hamid, M. F., Youssef, A., & Sharif, M. I. (2025). Knee Osteoarthritis Detection and Classification Using Autoencoders and Extreme Learning Machines. AI (Switzerland), 6(7). https://doi.org/10.3390/ai6070151

[5.] Panfilov, E., Saarakkala, S., Nieminen, M. T., & Tiulpin, A. (2025). End-to-End Prediction of Knee Osteoarthritis Progression With Multimodal Transformers. IEEE Journal of Biomedical and Health Informatics, 29(9), 6276–6286. https://doi.org/10.1109/JBHI.2025.3536170

[6.] Joseph, G. B., McCulloch, C. E., Nevitt, M. C., Lane, N. E., Majumdar, S., & Link, T. M. (2025). Machine learning models for clinical and structural knee osteoarthritis prediction: Recent advancements and future directions. Osteoarthritis and Cartilage Open, 7(3). https://doi.org/10.1016/j.ocarto.2025.100654

[7.] Sam Chandra Bose, A., Srinivasan, C., & Immaculate Joy, S. (2024). Optimized feature selection for enhanced accuracy in knee osteoarthritis detection and severity classification with machine learning. Biomedical Signal Processing and Control, 97. https://doi.org/10.1016/j.bspc.2024.106670

[8.] Bhateja, V., Dubey, Y., Maurya, N., Yadav, V. K., Shrivastava, S., Azar, A. T., Haider, Z., Amin, S. U., & Khan, Z. I. (2024). Ensemble CNN Model for Computer-Aided Knee Osteoarthritis Diagnosis. International Journal of Service Science, Management, Engineering, and Technology, 15(1). https://doi.org/10.4018/IJSSMET.349913

[9.] Yoon, J. S., Yon, C. J., Lee, D., Lee, J. J., Kang, C. H., Kang, S. B., Lee, N. K., & Chang, C. B. (2023). Assessment of a novel deep learning-based software developed for automatic feature extraction and grading of radiographic knee osteoarthritis. BMC Musculoskeletal Disorders, 24(1). https://doi.org/10.1186/s12891-023-06951-4

[10.] Ganesh Kumar, M., & Goswami, A. das. (2023). Automatic Classification of the Severity of Knee Osteoarthritis Using Enhanced Image Sharpening and CNN. Applied Sciences (Switzerland), 13(3). https://doi.org/10.3390/app13031658

[11.] Guan, B., Liu, F., Mizaian, A. H., Demehri, S., Samsonov, A., Guermazi, A., & Kijowski, R. (2022). Deep learning approach to predict pain progression in knee osteoarthritis. Skeletal Radiology, 51(2), 363–373. https://doi.org/10.1007/s00256-021-03773-0

[12.] Felfeliyan, B., Hareendranathan, A., Kuntze, G., Jaremko, J. L., & Ronsky, J. L. (n.d.). Title Improved-Mask R-CNN: Towards an Accurate Generic MSK MRI instance

segmentation platform (Data from the Osteoarthritis Initiative). https://doi.org/10.1016/j.compmedimag.2022.102056
[13.] B. Zhang, J. Tan, K. Cho, G. Chang, and C. M. Deniz, "Attention-based CNN for KL Grade Classification: Data from the Osteoarthritis Initiative," in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 2020, pp. 731-735. https://doi.org/10.1109/ISBI45749.2020.9098456
[14.] Bany Muhammad, M., Moinuddin, A., Lee, M. T. M., Zhang, Y., Abedi, V., Zand, R., & Yeasin, M. (2020). Deep ensemble network for quantification and severity assessment of knee osteoarthritis. Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019, 951–957. https://doi.org/10.1109/ICMLA.2019.00163
[15.] Ntakolia, C., Kokkotis, C., Moustakidis, S., & Tsaopoulos, D. (2020). A machine learning pipeline for predicting joint space narrowing in knee osteoarthritis patients. Proceedings - IEEE 20th International Conference on Bioinformatics and Bioengineering, BIBE 2020, 934–941. https://doi.org/10.1109/BIBE50027.2020.00158
[16.] Alexos, A., Kokkotis, C., Moustakidis, S., Papageorgiou, E., & Tsaopoulos, D. (2020, July 15). Prediction of pain in knee osteoarthritis patients using machine learning: Data from Osteoarthritis Initiative. 11th International Conference on Information, Intelligence, Systems and Applications, IISA 2020. https://doi.org/10.1109/IISA50023.2020.9284379
[17.] KOC_Net: Impact of the Synthetic Minority Over-Sampling Technique with Deep Learning Models for Classification of Knee Osteoarthritis Using Kellgren–Lawrence X-Ray Grade | MDPI
[18] shutterstock.com
[19] CNN Basic Architecture for Classification & Segmentation - Analytics Yogi
[20] https://data.mendeley.com/datasets/56rmx5bjcr/1