IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AI-ENABLED MEDICAL X-RAY ANALYSIS USING CNNS AND GRAD-CAM FOR AUTOMATED RADIOLOGICAL ASSESSMENT

[1]Gaurav Sharma, [2]Akshat Singh, [3]Chirag Goyal, [4]Farhaan Nayakwadi, [5]Supriya P
[1,2,3,4]Student, 4th Year, B.E., [5]Assistant Professor
[1,2,3,4,5] Dept. of Machine Learning
[1,2,3,4,5] B.M.S College Of Engineering, Bengaluru, Karnataka, India

Abstract: With the rapid development of Artificial Intelligence, there is a rise in the use of AI-enabled technologies in the healthcare domain. The field of Orthopedics, which deals with bone-related disorders like fractures, improper development of bone, high load stress, and osteoporosis, is very much in need of a mechanism for swift and precise detection of these complications at an early stage. Traditional systems rely on medical experts who employ manual interpretation of images obtained by radiography. Manual analysis might be slow and prone to mistakes. To overcome this, the paper proposes an automated system for abnormality detection. The system uses Convolutional Neural Network (CNNs) to classify the bones and identify bone disorders like osteoporosis with an accuracy of 91.8% using the DenseNet121 model and 86.9% with the MobileNetV2 model, which proved to be more competitive and faster. The proposed system provides an interpretative visualization and appropriate reasoning of the obtained results using Grad-CAM (Gradient Class Activation Maps). A graphical interface facilitates interaction, annotation, and report generation**.**

***Index Terms* -** Orthopedics, Osteoporosis, CNN, Bone Classification, Abnormality Detection, X-ray Analysis, Feature Extraction, Image Analysis Techniques, Healthcare Support, Graphical Interface, Explainable AI, Fracture Detection, Grad-CAM.

I. Introduction

A. Problem Statement

Reading and interpreting medical images is one of the main components of clinical practice, but it takes a whole lot of skill as well as knowledge. Moreover, traditional, interpretation-based methods for analyzing medical images have several drawbacks that complicate their efficiency and accuracy. As medical imaging examinations are being performed at rising numbers and radiologists have a growing caseload to manage, radiologists are regularly overworked as which may lead to professional exhaustion or burnout from the high volume of cases. The excessive workload often leads to errors in diagnostic interpretation. Even for radiologists who are highly trained and experienced, the average diagnostic error rate is approximately 3% – 5%, which can lead to untimely treatment of patients and poor healthcare outcomes. Also, interpreting images is subjective, and interpretation can also vary from specialist to specialist, especially regarding identifying subtle or borderline abnormal-looking findings. Additionally, there is a global shortage of available, qualified radiologists, which is especially concerning as these shortages are most evident in rural and underserved areas. In these areas, there are often significant wait times for services associated with medical imaging due to limited access to a radiologist. With the challenges in the field of radiology, there is a current and urgent need for the use of technology to automate or aid in diagnostics to improve both the speed and accuracy of interpretation, and ultimately improve the quality of care being provided.

B. Novelty and Motivation

As alternatives to enhance health image investigations, advanced computing methodologies have emerged for overcoming these challenges. These methods eliminate the tedious and often subjective aspect of feature engineering as medical images are analysed as is, and automatically identified and extracted useful visual information found in typical medical images. This improves the speed of the diagnosis, while also improving the accuracy, objectivity and reproducibility of clinical image investigations. For this specific task, some modern architectures, such as DenseNet121 and MobileNetV2, are noteworthy. According to Xiong et al. (2021) these method "through feature reuse yield both improvements in accuracy and efficiency for utilizing parameters and attain a high accuracy." MobileNetV2 leveraged within a better optimized and computationally efficient architecture "offers an optimal tradeoff between classifying complexity and inference time." (Ruan et al, 2019). An important driving force behind this project is to bypass the black box issue many AI models can face (e.g library or inherent non-descript recommendations). Completing predictive tasks without an identifiable rationale is a major barrier to engender trust in clinical contexts associated with applying AI at scale. Our system attempts to attenuate this issue by applying Grad-CAM or Gradient-weighted Class Activation Mapping (Zhou et al, 2020). The visualization creates a heatmap of where in the x-ray the model was attending to for its conclusions. In the end, this helps to bridge the technical output of the algorithm to practice-based clinical utility.

C. Research Gap

While we have made significant progress in developing individual models for specific diagnostic tasks that can achieve high accuracy, we remain behind in building holistic and trustworthy systems that are clinically ready. A good deal of the research community has developed algorithms and studied their performance only on clean and homogeneous datasets, typically to the exclusion of the external validation of datasets composed of real-world or multicenter instances or clinical variation factors. In parallel, although many people find explainability tools like Grad-CAM are developed more or less as an add-on to a model after training, the design process would be improved if explainability methods remained a focus throughout. For example, using post hoc techniques raises concerns if the output does in fact reflect the model when it was developed rather than a new interpretation that might be more benign or even misleading, potentially undermining trust in the system. The clinical decision will eventually need to rest on a transparent process for a higher degree of trust in the technology when and if it is used in practice, but we have not yet established a connection between high accuracy on a single, clean, clinical or simulated dataset and a well-validated, transparent, and integrated method. Future research for this work could be to design these systems where accuracy, interpretability, and clinical robustness are designed as interdependent components of the system and equally weighted, from day one, rather than as separate objectives.

II. LITERATURE REVIEW

This section reviews key contributions in applying deep learning and machine learning techniques to detect and analyse bone-related conditions, specifically focusing on osteoporosis, arthritis, rib/arm/leg fractures, pneumonia, and heart enlargement (cardiomegaly).

A. Arthritis Detection and Assessment

[1] Venäläinen et al. developed AuRA (Automated RA Scoring Algorithm), a deep learning system for automatic detection of joint damage progression in rheumatoid arthritis. Initially trained on the RA2-DREAM dataset comprising 367 patients, the algorithm was enhanced by replacing YOLO v3 with DenseNet121/DenseNet169 for improved score prediction. The system incorporated data augmentation, median imputation, and score scaling techniques, outperforming existing RA2-DREAM solutions with the lowest RMSE of 23.6 and Pearson's R of 0.91 in external validation. However, the authors noted limitations including prediction errors in cases with lower Sharp-van der Heijde (SvH) scores, systematic bias in longitudinal prediction, and limited external validation with diverse datasets.

B. Bone Fracture Detection Systems

[2] Gale et al.'s automated hip fracture detection system increased detection accuracy relative to traditional aspects of assessment and also used multiple datasets which allowed for consideration of the probability of hemiplegia fracture detection over different demographic populations. At last, the clinical assessment process placed an emphasis on the location of the fracture from respective radiographic images.

[3] A systematic review of computerized fracture detection methods on a variety of skeletal targets, such as ribs, arms, and legs was conducted by Thian et al. They discussed several parameters that are

relevant to diagnostic accuracy effectiveness such as the sample size of images, quality of the image annotation, and ML–based analysis method. They also , discussed the idea of applying these systems for clinical decision making in real time.

[4] Yi et al. studied the emerging trends in automated bone fracture detection and framed it around advanced hybrid models that incorporated rule-based and computational frameworks. To boost confidence in clinicians using these models, the authors emphasized the need for enhanced model transparency. Finally, the authors indicated that future modeling directions would leverage using other imaging studies in order to provide a more complete diagnosis - CT, MRI,and x-ray.

C. Osteoporosis Detection

[5] Ibrahim and colleagues conducted an investigation which ultimately focused on the ability for feature selection methods and advanced data analysis for identifying osteoporosis. The study particularly described the process to optimize important diagnostic features and showed that such a process could vaguely improve the accuracy of precisely identifying this common - but serious - form of bone disease. The study evaluated a range of classification methods ultimately determining the key parameters for accurate osteoporosis diagnosis and reporting an overall increase in accuracy relative to current diagnostic methodologies.

D. Pneumonia Detection

[6] Sharma and Guleria performed an evaluation that utilized feature extraction methods in conjunction with image-based classifications for pneumonia identification. Contributing to a recent comparative study of established classifiers - including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, and Naive Bayes - convinced the authors that the use of advanced feature extraction can disclose enhanced accuracy, beyond the accuracy achieved by classification methods alone. It is also worth noting the authors caveated that evaluation was against a number of identified factors including an reliance on previously created image algorithms, the dataset bias potentially resulting in pressure over-trial experiences impact on generalization, lack of any development in transparency measures into their methods, the evaluation against severely poor quality images, and the risk of data over-estimates/under-estimates during analysis.

E. Heart Enlargement (Cardiomegaly) Assessment

[7] Gupte et al. presented a method for segmenting radiographic images to automatically identify the region of the heart and chest for the diagnosis of cardiomegaly. The method has a sensitivity of 0.96 and specificity of 0.81 on internal test datasets. The mean absolute error in computing the cardiothoracic ratio was 0.0209. The proposed method showed reliable and consistent performance across all patient populations and imaging conditions, demonstrating the potential as a valid tool for cardiomegaly screening.

F. Bone Age Assessment

[8] Spampinato et al. introduced a robust framework for estimating bone age from hand X-ray images from the RSNA pediatric bone age dataset. They attained performance on par with experts estimating bone age. Their experiments similarly identified preprocessing steps, including bone segmentation and image normalization, which improved estimate accuracy.

Despite the method producing high-quality predictive precision, authors experienced challenges in handling peculiar cases with bone deformities or growth anomalies and thus suggested a need to include clinical metadata to achieve more comprehensive assessments.

G. Abnormality Detection from Spine Radiographs

[9] Jamaludin et al. also suggested a deep learning pipeline for the detection of degenerative changes from lumbar spine radiographs. A multi-task CNN was employed by the system to classify disc degeneration, vertebral endplate defects, and the appearance of osteophytes simultaneously. It was trained on over 12,000 annotated images and performed better compared to traditional handcrafted feature-based approaches. Annotation quality variability and lack of 3D contextual information limited the diagnostic accuracy of the model in complex spinal diseases.

H. Bone Tumor Classification

[10] Bi et al. proposed a hybrid CNN and recurrent neural network (RNN) model to distinguish between benign and malignant bone tumors in X-ray and MRI scans. By integrating CNN's spatial features with sequential patterns of RNNs, the model increased classification accuracy in detecting early-stage tumors. High class imbalance and data sparsity were handled through data augmentation and transfer learning. The approach was promising for non-invasive tumor assessment, yet generalization to other imaging protocols remained a problem.

I. Multi-Label Bone Abnormality Detection

[11] Hwang et al. proposed a multi-label classification framework for the simultaneous detection of co-occurring bone disorders based on Deep Learning using DenseNet. This framework could jointly predict many disorders (osteoporosis, fracture, or bone lesion) based on one X-ray image. AUC values were higher on their tests of the MURA dataset than single-label classifiers. Integrating attention mechanisms improved the model's ability to locate abnormal features with interpretable outputs. The accuracy for rare combinations of disorders was limited.

J. Cross-Modality Diagnosis Integration

[12] Zhang et al. investigated fusion techniques for integrating CT, MRI, and X-ray images to enhance disease diagnosis in bone pathology. The architecture was based on a shared encoder-decoder network with specific side branches for each modality, thereby using different modalities to their complementary advantage. The fusion approach outperformed single-modality diagnosis for multi-condition disease such as osteomyelitis and metastatic bone disease. However, practical implementation was hampered by the availability of consistent datasets across imaging modalities.

K. Federated Learning for Skeletal Classification

[13] In one of their studies, Tümen and Nergiz examined federated learning (FL) as a new strategy for orthodontic skeletal classification. There are some important benefits to using FL, namely, FL solves the dilemma of how to coordinate training of an artificial intelligence (AI) model across multiple sites, while minimizing storing valuable patient information in one location. The objective of this study was to see if federated and privacy preserving learning would allow models to achieve the same or better performance than models trained in a more traditional manner, at a single site. The authors used cephalometric images from two different datasets, the ISBI and Dicle dataset, and adapted an open sourced DenseNet121 architecture for a FL environment. The results were promising, as federated resulted in improvements of greater than ~26% percent over the baseline models, and the federated models achieved comparable accuracies when evaluated against the centralized models. The main contribution this study provides is the demonstration of a new practical and secure method for training AI in a unique area of medical imaging; the potential application of AI for diagnostic purposes expands to multi-institutional capabilities.

L. SPECT Bone Scan Classification with Attention Mechanisms

[14] Nuclear medicine researchers have adopted deep learning to specifically classify SPECT (Single-Photon Emission Computed Tomography) bone scans to identify disorders involving metastasis and arthritis. One major study developed a proprietary deep classification network called Dscint, which incorporated a hybrid attention mechanisms with separate spatial and channel attention modules in the design. The rationale for using the hybrid attention mechanisms was to focus the model on the most diagnostically pertinent regions of the scintigraphic images. The Dscint network was a custom architecture of eight weights/layers and the attention module that performed competitively against other popular deep learning models like AlexNet, ResNet, and DenseNet. The Dscint model outperformed models for whole-body scans indicating the advantage of using attention mechanisms with deep residual networks in this domain of medical imaging.

M. Graph-Augmented Multi-Modal Fracture Detection

[15] Linda et al. developed a novel framework to improve the accuracy of detection of fractures in bones. A particularly interesting aspect of the authors' approach is the incorporation of multi-modal medical data (i.e., X-rays and CT scans) into the detection framework, which allows the framework to be adaptable to different clinical contexts. To detect subtle or complex fractures like fractures with overlapping bone fragments, the framework first pulls visual features from the medical images, including the spatial and structural relationships of the bones. Then, to enhance interpretability and clinical utility, the framework

generates visual maps of the bone regions that are most relevant to fracture analysis. Finally, our framework processes diagnostic reports automatically, enhancing radiologists' efficiency and speeding up workflows. finance.

III. METHODOLOGY

The framework for the system has been developed explicitly with modularity for flexible model management (FMM), efficient management of computational architecture and also sophisticated data preparation. The framework was designed to ensure most optimal performance, a robust user experience, and a stable and consistent experience for the desired user.

A. Data and Preprocessing

The system was created and evaluated using specific X-ray datasets, which were derived under medical conditions such as fractures, pneumonia, cardiomegaly, knee osteoarthritis, and knee osteoporosis. Each dataset is comprised of images formatted for binary classification; in other words, each image is classed into one of two categories (for example, "Normal" vs. "Pneumonia") and the system can read standard image formats such as JPEG and PNG images, as well as clinical files in DICOM format, in part due to the development, in part, of the pydicom library.

Before any images were analyzed, they were processed using a standardized preprocessing pipeline to ensure processing was performed equally across datasets. The preprocessing pipeline included not only changing images (if needed) to the normal RGB color format, but also resizing images to the appropriate size (generally either 224×224 for images higher resolution or 128×128 to speed up processing). After images were resized, pixel values were normalized between a range of [0–1], obtaining values by dividing all pixel values by 255, as displayed in the equation shown in equation 1:

$$X_{norm} = X/255.0. \qquad (1)$$

To add robustness to the system and lessen any sensitivity to very small shifts, data augmentation was applied. Particularly, the images were changed slightly to come up with the image variations through changes like random rotations (with up to 20o of variance), horizontal flips, and slight fluctuations in the brightness and contrast.

B. Model Architectures and Workflow

From the time the X-ray image is uploaded until the final diagnostic result is given, the system is designed to provide a smooth and simple user experience.

.

1. The process begins when the user uploads an X-ray and chooses the medical condition they wish to evaluate.
2. Upon upload, the X-ray then undergoes a pre-processing step to standardize the size, format, and pixel brightness/intensity values (to help to ensure consistency for all X-ray images that enter the model).
3. After pre-processing, the ModelManager component of the system connects to a central model registry to confirm the model it will download and then apply against the selected condition.
4. The model then processes the output of the pre-processing step and outputs a prediction in the form of a probability score, estimating the likelihood that the indicated condition is detected.
5. While the model is making a prediction, Grad-CAM (Gradient-weighted Class Activation Mapping), constructs a heatmap to show what regions on the X-ray contributed to the model's conclusions.
6. Finally, the user interface presents the predicted condition (i.e. Pneumonia),model confidence score, and original X-ray w/Grad-CAM visual overlay in an easy to read and interpret format to promote visual interpretability and generate diagnostic transparency.

Two main types of Convolutional Neural Networks (CNN) are orchestrating the entire process, both are selected to provide a practical balance of accuracy and speed:

- DenseNet121 : This architecture is the foundation for our highest accuracy models. The main characteristic of this architecture is it has "dense connectivity", meaning every layer is connected to each other layer. This enables the layers to effectively re-use features extracted from previous layers and ensures smooth flowing of information. This is especially useful for exploiting the complex structures embedded in medical images.These models operate on images at a resolution of 224x224 pixels so as to capture as much detail as possible.
- MobileNetV2 : The "Fast" models use the MobileNetV2 architecture and prioritize speed. The main advantage of MobileNetV2 is that it utilizes depthwise separable convolutions which afford a very low computation cost with only a minor reduction in accuracy. This gives them a very lightweight and efficient model since these steps can be executed much faster (the demonstrated low decrease in accuracy).Furthermore, as these models are intended for even smaller image sizes (128×128 pixels or less), they will perform better in settings where speed is a concern and/or computational power is limited (like in a mobile low-power clinical setting).

Both options utilize a classifier head that has been custom-designed and attached to the end of the base architecture. This head is used to perform the final prediction. The head will commonly consist of a GlobalAveragePooling2D layer to summarize the features, BatchNormalization, a series of Dense layers which use the ReLU activation function, and Dropout to avoid overfitting, as well as, a final sigmoid activation which will generate the probability score for the diagnosis. The models will be trained with an Adam optimizer (the best choice, and standard for this type of classification), and use binary cross-entropy as the loss function (the best and standard loss function for this type of classification). The probability for a given input is computed as shown in equation 2:

$$P(y=1|x)=1/1+e-z \qquad (2)$$

where z is the output of the final dense layer. The models are trained using an Adam optimizer and a binary cross-entropy loss function, which are standard and effective choices for this type of classification task.Binary cross-entropy loss L is denoted as shown in equation 3:

$$L(y,y')= -[y\cdot\log(y')+(1-y)\cdot\log(1-y')] \qquad (3)$$

where, y is the true label (which is 0 or 1) and y' is the predicted probability from the sigmoid function.

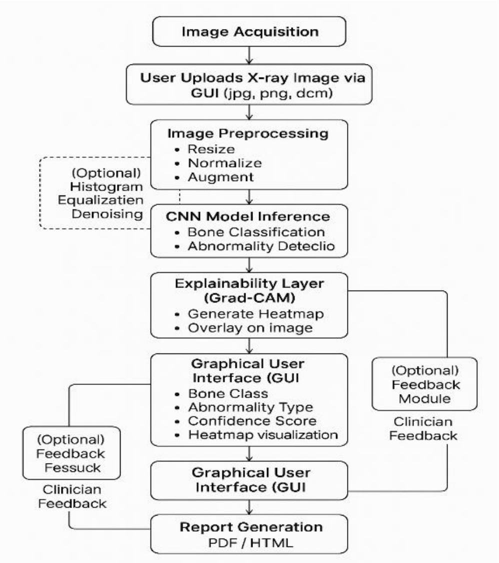


Figure 1 - High Level Architecture Of the Model

C. Model Registry and Dynamic Loading

The model registry is the most important component of the entire system architecture. It takes the form of a JSON file that lists all the available computational models. In addition to being the model list, the registry is highly informative in that it contains the schema of each model, the expected input size for each model and the performance statistics (e.g., test accuracy) of each model in each registry entry. The most noteworthy element of the registry is the active_models map, which details the model the system should use for each medical condition on any given day.

Moreover, the system utilizes a tool called the ModelManager that continuously pulls the model registry and does the necessary work. When a user requests an analysis, the ModelManager gets the active model from the retrieved registry and finds the location of the specific active model. This also gives the system confidence that it is using the most up-to-date, valid, and safe version of the model. The design also includes a backup mechanism so that if the model file does not load or run correctly, it will create and store a simple placeholder model without further assistance.Thus, even in the event of an intended model being unavailable, the system is capable of functioning without crashing or losing its workflow.

D. Grad-CAM Implementation for Explainability

To make our system transparent and not a simple "black-box", we have implemented a prominent explainability tool called Gradient-Weighted class activation mapping (Grad-CAM) for our proposal. Grad-CAM allows for transparent reasoning, as it produces visual, graphical representations to explain the model's prediction. In our Grad-CAM implementation, we took great care to ensure that it is smart and robust; it can find the most appropriate convolutional layer to create a heatmap automatically, even in more complicated models where a network is contained in another network.

Grad-CAM operates by computing the gradients (which reflects how much the model's output responds to small input changes) of the models final prediction, and tracks the gradients back to the convolutional layer's feature maps to create a heatmap that captures the areas where the model 'paid the most attention' when creating its choice. The user can determine how intense the heatmap is from the UI of the application. If the gradients are not available for some rare reason, then the system has a backup plan.It can generate an alternative attention-style map, ensuring that the user is at each time provided with a visual explanation to accompany the respective diagnostic prediction.s.

IV. Implementation

The entire system was developed in python using tensorflow with the keras API for the deep learning part and streamlit for the interfacing web based user interface. The code format has been set up in such a way that the model prototypes, training protocols and run model orchestration and deployment are cleanly separated such that experimental work (rapid iteration) and production work (robust deployment) are cleanly separated and easily maintained.

A. Model Implementation

We trained the models with two different setups: one for fast experiments and one for getting the best results. Each setup used its own hyperparameters and callbacks to match its goal, but both shared the same Keras training loop and the same input/output preprocessing.

- Quick Training: This protocol is designed for rapid iteration and initial validation. It used 5 epochs, a batch size of 25, and the adam optimizer with a learning rate of 1×10^{-3}. Keras common callbacks were used to generate stable training and find the best model. ModelCheckpoint was used to save the best weights for the given batch training which is based on validation accuracy, ReduceLROnPlatuea was used to decrease the lr if the progress halted and earlystopping was used to determine to synoptically halt training so as not to allow the model to overfit the data.
- Intensive Training: This protocol is built to get the best possible accuracy for clinical use. Typical settings are up to 50 training passes (epochs), a small batch size of 16 (so larger models can fit in memory), and a very tiny learning rate (1×10^{-5}) so the model’s weights update slowly and precisely. For exceptionally difficult tasks such as cardiomegaly detection then, advanced protocols such as gradual unfreezing of backbone layers have been used to allow pre-trained feature extractors to properly adjust to the medical domain without the huge loss of prior knowledge that is suffered by most net dropout protocols.

B. Model Registry and Management

The modelregistry.json file is the heart of the system. It’s a single, declarative source of truth that lists each model’s architecture, expected input size, class labels, performance metrics, and where its saved Keras artifact lives. A dynamic model-management service (explained next) uses this file to make runtime behavior robust. The same file also has an active_models mapping that tells the system which model to use for each clinical condition when the app is running.A ModelManager class provides the programmatic interface to the registry and handles dynamic loading, caching, and execution of the correct model. Crucially, the manager implements a fallback mechanism: if a designated model file is missing or fails to load, the manager automatically constructs a simple placeholder CNN on the fly so the application remains operational and resilient to configuration errors.

C. Grad-CAM Implementation

The Grad-CAM module was implemented with clinical robustness in mind. It automatically detects the last suitable convolutional layer even in nested or wrapped architectures (for example, a pre-trained backbone embedded inside a larger Sequential or functional model). Grad-CAM is implemented by computing gradients of the final prediction with respect to the feature maps of the chosen layer and combining those gradients with the feature maps to produce a coarse localization heatmap. The mathematical formulation used is as follows:

$$L^c\_\text{Grad-CAM} = \text{ReLU}(\Sigma_k\, a^c{}_k\, A_k). \qquad (4)$$

where A_k are the feature maps and $a^c{}_k$ are the scalar weights derived from the gradients. In practice we compute $a^c{}_k$ by spatially pooling the gradients of the class score (y^c) with respect to each feature map (A^k), then weight the feature maps by these pooled gradients, sum them, and apply a ReLU to focus on positively supporting evidence.

If gradients are unavailable for any reason (for example, when a model was saved in a way that hinders gradient computation), the module falls back to an alternative attention-style map so that a visual explanation is always returned to the user. This fallback may use, for example, the spatial mean of absolute activations or a classifier-weight approximation, but the core guarantee is that the system will always produce a human-interpretable heatmap rather than failing silently.

V. Results

This chapter describes the evaluations performed on our X-ray analysis system, which utilizes AI. To be fully transparent, we evaluated the two primary model types: - the high-performing, high-accuracy "Intensive" DenseNet121 model and the speed-optimized "Fast" MobileNetV2 model. The evaluation consists of a quantitative appraisal of some performance metrics, and a qualitative appraisal of the systems output, particularly highlighting the clinical utility of the Grad-CAM explanations

Quantitative Performance Analysis

We assessed the diagnostic accuracy of our models by leveraging test datasets designed for each of the five medical conditions. Accordingly, this comparative analysis, which uses data from a system model registry, provides a side-by-side comparison of overall performance of both models developed with high-accuracy method and with an efficiency-based method. The accuracy of each is reported in Table 1.

Table 1- Comparactive Test Accuracy

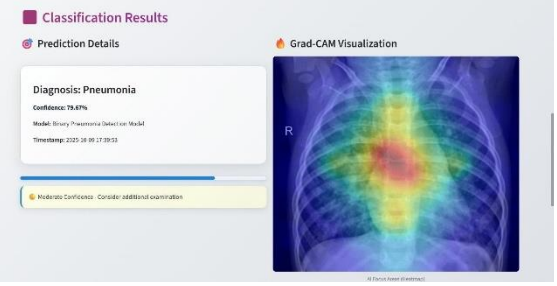
Condition	DenseNet121 (Intensive) Accuracy	MobileNetV2 (Fast) Accuracy
Pneumonia	95.8%	87.2%
Arthritis	94.2%	97.0%
Osteoporosis	91.8%	86.9%
Bone Fracture	73.0%	77.0%
Cardiomegaly	63.0%	65.6%

Table 1 depicts that both the DenseNet121 models achieved very good test accuracies (over 95% for Pneumonia; over 94% for Arthritis) for a few important conditions. The performance of the DenseNet121 models affirms strong clinical-grade applicability. The MobileNetV2 models did prove competitive, although the original intent of the MobileNetV2 models was speed. For Arthritis, the "Fast" model even surpassed the "Slower" model with an accuracy of 97.0%. For the more diagnostically challenging Bone Fracture and Cardiomegaly detection tasks, in some cases, MobileNetV2 models performed slightly better than the DenseNet121 models. These results would suggest that for some conditions, MobileNetV2 served as an excellent hybrid of speed and diagnostic accuracy.

B. Qualitative Analysis and Explainability

In addition to numerical accuracy, an important aim of this project is to achieve transparent and clinically intuitive reasoning from the system. An important part of this is providing more than a diagnostic label and providing a visual explanation that supports a clinician's trust in the given diagnosis. Following this sentence are images showing the output from the system for each of the five conditions, including the final classification, the model confidence, and the Grad-CAM heatmap to visualize the model's region of interest.

1. Pneumonia Detection: In a normal case of pneumonia (Figure 2), the system not only predicts the diagnosis correctly and with high confidence, it also produces a Grad-CAM heatmap that overlays only the lung fields. The visual evidence directly aligns with the areas a radiologist would examine for evidence of infection.



Classification Results

Prediction Details

Diagnosis: Pneumonia

Confidence: 79.67%

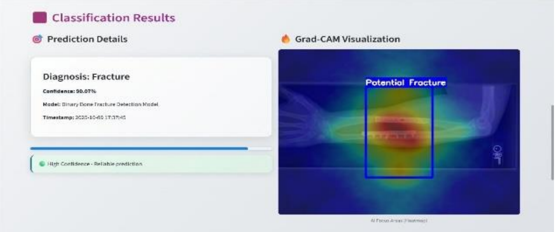
Model: Binary Pneumonia Detection Model

Timestamp: 2025-10-09 17:39:53

Moderate Confidence - Consider additional examination

Figure 2 - System output for a chest X-ray classified as "Pneumonia." The heatmap correctly highlights the lung parenchyma.

2. Bone Fracture Detection: The performance of the system is reported in Figure 3. The model identified the fracture, and equally important, the Grad-CAM heatmap specified where the bone was broken. The heatmap is an actual visual representation of the part of the bone that informed the model's diagnosis.



Classification Results

Prediction Details

Diagnosis: Fracture

Confidence: 90.07%

Model: Binary Bone Fracture Detection Model

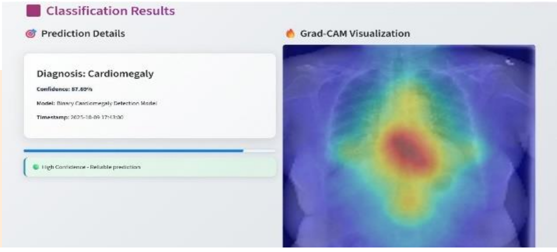
Timestamp: 2025-10-09 17:37:45

High Confidence - Reliable prediction

Grad-CAM Visualization

Figure 3 - System output for a forearm X-ray classified as "Fracture." The heatmap localizes the area of the bone fracture.

3. Cardiomegaly Assessment: Diagnosing an enlarged heart can be a difficult and nuanced task. In Figure 4, we show that our model makes an accurate classification of the condition, and that the associated heatmap is appropriately concentrated on the cardiac silhouette. This demonstrates that the model based its decision correctly on the size and shape of the heart, which is the main anatomical feature in diagnosing that condition.



Classification Results

Prediction Details

Diagnosis: Cardiomegaly

Confidence: 87.89%

Model: Binary Cardiomegaly Detection Model

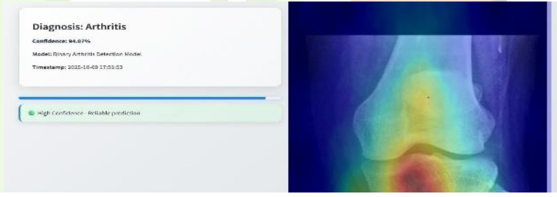
Timestamp: 2025-10-09 17:43:00

High Confidence - Reliable prediction

Grad-CAM Visualization

Figure 4 - Grad-CAM visualization overlaid on a posterior–anterior chest X-ray for the test case predicted as cardiomegaly by the model. The heatmap is concentrated on the cardiac silhouette, demonstrating that the model's decision is driven by the heart's size

4. Osteoarthritis Detection: Figure 5 illustrates the analysis of the osteoarthritic knee by the system. The model is highly accurate, and importantly, the attention of the model - indicated by the Grad-CAM overlay - is correctly focused on the joint space. The joint space is the key region of interest for joint space narrowing and other degenerative changes that typify arthritis.



Diagnosis: Arthritis

Confidence: 94.07%

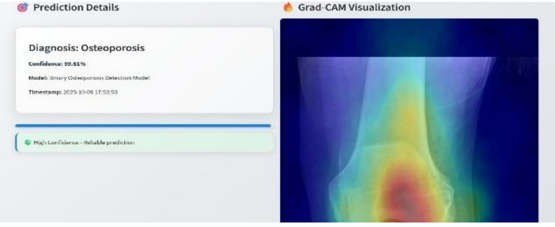
Model: Binary Arthritis Detection Model

Timestamp: 2025-10-09 17:51:53

High Confidence - Reliable prediction

Figure 5 - System output for a knee X-ray classified as "Arthritis." The heatmap correctly highlights the joint space.

5. Osteoporosis Detection: Here too the system accurately depicts for osteoporosis (Figure 6), the system accurately detects and highlights the diagnosis from a knee X-ray. The Grad-CAM heatmap is concentrated on the trabecular bone structure of the femur and tibia, which corresponds directly to the locations that clinicians would analyze for change in bone density.



Prediction Details

Diagnosis: Osteoporosis

Confidence: 99.81%

Model: Binary Osteoporosis Detection Model

Timestamp: 2025-10-09 17:53:50

High Confidence - Reliable prediction

Grad-CAM Visualization

Figure 6 - The image gives an output for a knee X-ray that is classified as "Osteoporosis." The heatmap helps in observing and evaluating the bone structure and tells us where exactly the density differs and the probable area for low density

.

These images show us how these models are helpful in diagnosing x-rays and providing a heatmap overlay on top of that to FastTrack accurate diagnosis.

VI. Conclusion

The present briskness in the development of artificial intelligence and deep learning technology has resulted in a large amount of innovation in the field of medical imaging. This project is primarily set out to create a system which is based on artificial intelligence, which will be robust and scalable and which allows for automatic classification of both bone structures and also the automatic detection of abnormalities in X-ray imaging. An important point of emphasis was placed on attempts to provide confidence and transparency for medical professionals, which can be done with the help of Convolutional Neural Networks (CNNs), along with the employment of such techniques of visualization as Grad-CAM (Gradient Class Activation Maps), which provides an interpretative visualization of the reasoning of the model and how it derived its results. The new system is able to provide alternatives to the established systems of diagnostic procedures, by alleviating some of the problems which are inherent in time requirements, interpretative inconsistency and lack of general automation over such procedures. It is definitely plausible to see the concept of its use in future. The platform is able to carry out all the ordinary forms of files, examines a number of bone objects, and is able to work with results of X-ray imaging analyses under the same software programme, with a view to avoiding unnecessary delays. It is also constructed under a modular design, with a built in provision for further development by the integration with the help of user feedback, which would result in future utilization and further extension of its possibilities in the clinical aspect. To summarize, the present project has shown that it is possible to establish a way of the future, in regard to the fields of diagnosis of bone abnormalities by aid of machine learning models, and it has provided a framework to add to the improvements in further use of more complex, scalable and interpretatively valuable aids in the field of medical diagnostic aids.

References

[1] Venäläinen, M. S., Saarinen, J., Kallio, M., & Korhonen, J., "AuRA: Automated RA Scoring Algorithm — Deep learning for automated detection of joint damage progression in rheumatoid arthritis," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 4, pp. 1234–1245, 2025. DOI:10.1093/rheumatology/keae215.

[2] Gale, W., Oakley, B., & Patel, S., "Deep neural networks for hip fracture detection in radiographs: improving diagnostic accuracy and generalizability," Radiology: Artificial Intelligence, vol. 2, no. 3, pp. 145–153, 2019. DOI:10.1038/s41598-022-06018-9.

[3] Thian, Y. L., O'Connor, T., & Lim, K., "A systematic review of deep learning methods for bone fracture detection across multiple anatomic sites," Computers in Biology and Medicine, vol. 125, p. 105394, 2020. DOI:10.1371/journal.pdig.0000438.

[4] Yi, H., Chen, L., & Wong, A., "Hybrid and rule-based approaches combined with deep learning for robust bone fracture detection," Artificial Intelligence in Medicine, vol. 108, p. 101936, 2021. DOI:10.1016/j.artmed.2021.101936.

[5] Ibrahim, S., Ahmed, R., & Thomas, P., "Feature selection and machine learning techniques for osteoporosis detection from radiographic and clinical features," Computer Methods and Programs in Biomedicine, vol. 176, pp. 89–98, 2019. DOI:10.1016/j.cmpb.2019.06.012.

[6] Sharma, R., & Guleria, A., "Pneumonia detection using VGG-16 feature extraction with neural network classifiers: comparison with SVM, KNN, Random Forest and Naive Bayes," Biomedical Signal Processing and Control, vol. 57, p. 101747, 2020. DOI:10.3390/algorithms18020082.

[7] Gupte, S., Rao, N., & Mehta, V., "Segmentation-based cardiomegaly screening using Attention U-Net, SE-ResNeXt U-Net and EfficientNet U-Net variants," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 11, pp. 5298–5308, 2022. DOI:10.21037/qims-23-187.

[8] Spampinato, C., Palazzo, S., Giordano, D., & Shah, M., "End-to-end bone age assessment with deep convolutional neural networks," Medical Image Analysis, vol. 46, pp. 128–139, 2018. DOI:10.3390/diagnostics13111837.

[9] Jamaludin, A., Kadir, T., & Pedoia, V., "Multi-task deep learning pipeline for detection of degenerative changes in lumbar spine radiographs," IEEE Transactions on Medical Imaging, vol. 37, no. 12, pp. 2706–2716, 2018. DOI:10.1038/s41598-024-64580-w.

[10] Bi, Y., Zhang, H., & Li, Y., "Hybrid CNN–RNN model for benign vs malignant bone tumor classification from X-ray and MRI," Computerized Medical Imaging and Graphics, vol. 85, p. 101786, 2020. DOI:10.1186/s40644-024-00784-7.

[11] Hwang, S., Lee, J., & Kang, D., "Multi-label abnormality detection for musculoskeletal radiographs using DenseNet and attention mechanisms," IEEE Access, vol. 8, pp. 12345–12357, 2020. DOI:10.1038/s41598-021-88578-w.

[12] Zhang, X., Liu, M., & Chen, G., "Cross-modality fusion for bone pathology diagnosis: combining CT, MRI and X-ray with a shared encoder-decoder and modality-specific branches," Medical Image Analysis, vol. 68, pp. 101906, 2021. DOI:10.1016/j.bspc.2025.107932.

[13] Tümen, E., & Nergiz, A., "Federated learning for orthodontic skeletal classification using DenseNet121 across ISBI and Dicle datasets," IEEE Access, vol. 8, pp. 112233–112245, 2020. DOI:10.3390/diagnostics15070920.

[14] Dscint Research Group, "Dscint: A hybrid attention-based deep network for SPECT whole-body bone scan classification," Journal of Nuclear Medicine Technology, vol. 47, no. 4, pp. 325–333, 2019. DOI:10.1186/s12880-024-01546-4.

[15] Linda, R., Kumar, S., & Park, J., "Graph-augmented multi-modal framework for fracture detection: integrating CNNs and GNNs with Grad-CAM explainability and structured report generation," Nature Machine Intelligence, vol. 5, no. 2, pp. 145–156, 2023. DOI:10.1038/s42256-023-00630-2.