INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

Automated Prediction System for Bone Cancer Detection and Bone Age Assessment using Deep Learning Models

Harshit Vora[1], **Seema Mahajan**[1], **Yogesh Kumar**[2]*

1 CE, Indus University, Rancharda- Ahmedabad, India
2 CSE, School of Technology, Pandit Deendayal Energy University, Gujarat, India

OPEN ACCESS

Received: 16/07/2025
Accepted: 14/08/2025
Published: 08/09/2025

Citation: Vora H, Mahajan S, Kumar Y (2025) Automated Prediction System for Bone Cancer Detection and Bone Age Assessment using Deep Learning Models. Indian Journal of Science and Technology 18(33): 2701-2714. https://doi.org/10.17485/IJST/v18i33.1245

* **Corresponding author**.

Yogesh.kumar@sot.pdpu.ac.in; Yogesh.arora10744@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objective: To develop an automated system for accurate bone age assessment and bone cancer detection in paediatric and oncological care, addressing the limitations of traditional manual X-ray evaluation methods, which are often subjective and inefficient. **Methods:** This study uses a deep transfer learning-based framework trained on a dataset of annotated bone images, including both malignant and non-cancerous cases. Advanced image preprocessing, including segmentation using contour detection and feature extraction techniques, is applied to enhance input quality and model performance. **Findings:** Among the evaluated models, EfficientNetB0 achieves the highest accuracy for bone age prediction with a root mean squared error (RMSE) of 0.1637. For bone cancer classification, DenseNet201 and InceptionResNetV2 exhibits superior performance, achieving precision and recall scores of 0.985 and 0.945, respectively. These results indicate strong model generalization and diagnostic capability on unseen medical data. **Novelty:** This work demonstrates a robust application of fine-tuned deep transfer learning models integrated with image segmentation and feature enhancement for dual-purpose prediction of bone age estimation and cancer detection. The dual-task architecture and use of a large, annotated dataset make this system a scalable solution for clinical integration, potentially reducing diagnostic variability and supporting faster decision-making in medical practice.

Keywords: Artificial Intelligence; Medical Imaging; Bone age; Bone cancer; Deep Transfer Learning; Paediatricians

1 Introduction

Bone age is defined as the biological age of an individual's skeletal system which serves as a valuable indicator in evaluating growth, development, and potential health concerns, particularly in the paediatric population. The intricate relationship between bone diseases and bone age raises complex questions about how pathological conditions can influence the normal trajectory of skeletal development [(1)]. One such common bone disease is bone cancer which can manifest in any bone across the body, although it is

more frequently observed in the long bones such as the legs and arms. Primary bone cancers are those that develop within the bone. Many cancers that start in organs or other regions of the body may expand to the bones and other body components. These growths are known as metastatic or secondary bone malignancies [2]. Bone tumors are uncommon. They account for less than 1% of malignancies in the United States. While they can occur at any age, they are more prevalent in kids, teens, and young adults rather than older persons. They are divided into four stages i.e. low grade, high grade, spread within the same bone and spread to the other areas [3].

In this era, latest breakthroughs in artificial intelligence (AI) can tackle these issues than with powerful learning models and better data analysis. AI techniques present a good opportunity in the early diagnosis of bone cancer using medical images that small differences in patterns may depict normally but are noticeable by an AI system and reduce diagnostic. Several researchers have explored diverse methodologies for bone disease detection and age prediction, yet various limitations persist. Xinyang et al. (2024) [4] developed a radiomics-based machine learning model using bpMRI for predicting bone metastasis in prostate cancer patients, achieving 82.8% accuracy; however, their study was constrained by a small dataset and limited scope to prostate-specific cases. Jeong et al. (2024) [5] designed a deep learning model to detect nasal bone fractures, demonstrating perfect sensitivity, but the study was in its preliminary phase and lacked validation across other fracture types. Gawade et al. (2023) [6] applied CNN and SVM models for osteosarcoma classification, yet their approach was limited to a single cancer type without incorporating segmentation for enhanced precision. Another study utilized edge detection with histogram of oriented gradients (HOG) and traditional classifiers like SVM and Random Forest to differentiate healthy and cancerous bones, but it lacked deep learning-based feature extraction and multiclass capability. Sahin (2023) [7] relied on classical machine learning and image processing methods for fracture detection, achieving high accuracy, though the absence of deep learning restricted model generalization. Anand et al. (2022) [8] introduced a DC-ELM model for histopathological cancer classification with high accuracy, yet the reliance on hand-crafted features limited its adaptability. Suganeshwari et al. (2023) [9] used transfer learning with mutual information-based feature selection and SVM yet relied on static feature extraction. Ramasamy et al. (2023) [10] implemented a hybrid CNN approach with adaptive algorithms, achieving 99.45% accuracy, although their method was not evaluated in clinical settings. Choi et al. (2024) [11] investigated bone age prediction using contrast-enhanced CNNs, reporting significant improvements with histogram equalization, yet the models struggled with generalization. Kassem et al. (2023) [12] designed an explainable AI system for pelvis fracture detection, achieving high accuracy, though the study was constrained to a narrow diagnostic category. Overall, while these studies contribute significantly to the field, they often suffer from limitations such as small datasets, narrow scope, lack of integration between segmentation and classification, or absence of real-time clinical validation.

Previous research on bone age estimation and bone cancer detection has explored both classical and deep learning approaches. While deep learning methods, particularly CNN-based architectures [5,6,8–12], have shown improved performance, most studies focus on a single diagnosis, such as osteosarcoma or fracture detection, or on a single task, either classification or regression. Common shortcomings include small datasets that limit model robustness, the absence of segmentation steps that results in noisy feature learning, a narrow scope addressing only one task, and poor external validation, which restricts clinical applicability.

In contrast, our work introduces a dual-task system capable of both bone cancer detection and bone age estimation within a single deep learning framework. By combining advanced preprocessing, segmentation, and transfer learning on a large, annotated dataset (12,611 images), the proposed model enhances feature quality, improves generalization, and addresses the task integration gap in prior studies.

2 Methodology

In this section, the process that has been used in developing the deep learning model for detecting and predicting bone cancer along with the bone age has been discussed, as depicted in Figure 1.

2.1 Dataset description

The GRAZPEDWRI-DX dataset is a curated collection of pediatric wrist trauma radiographs, with the name derived from "Graz," "Pediatric," "Wrist," and "Digital X-ray." It contains de-identified images in PNG format, publicly available for research. The dataset includes radiographic studies from 6,091 pediatric patients aged 0.2 to 19 years (mean age: 10.9 years), comprising 2,688 females, 3,402 males, and one patient with unknown gender. Initially stored in DICOM format commonly used in clinical environments for its inclusion of both image and metadata, the images were converted to PNG due to practical limitations of using DICOM outside clinical settings. Since DICOM typically stores images in 12-bit or 16-bit grayscale, the pixel values were normalized to 16-bit before conversion using the "cv2" module, ensuring grayscale fidelity was maintained in PNG format. The

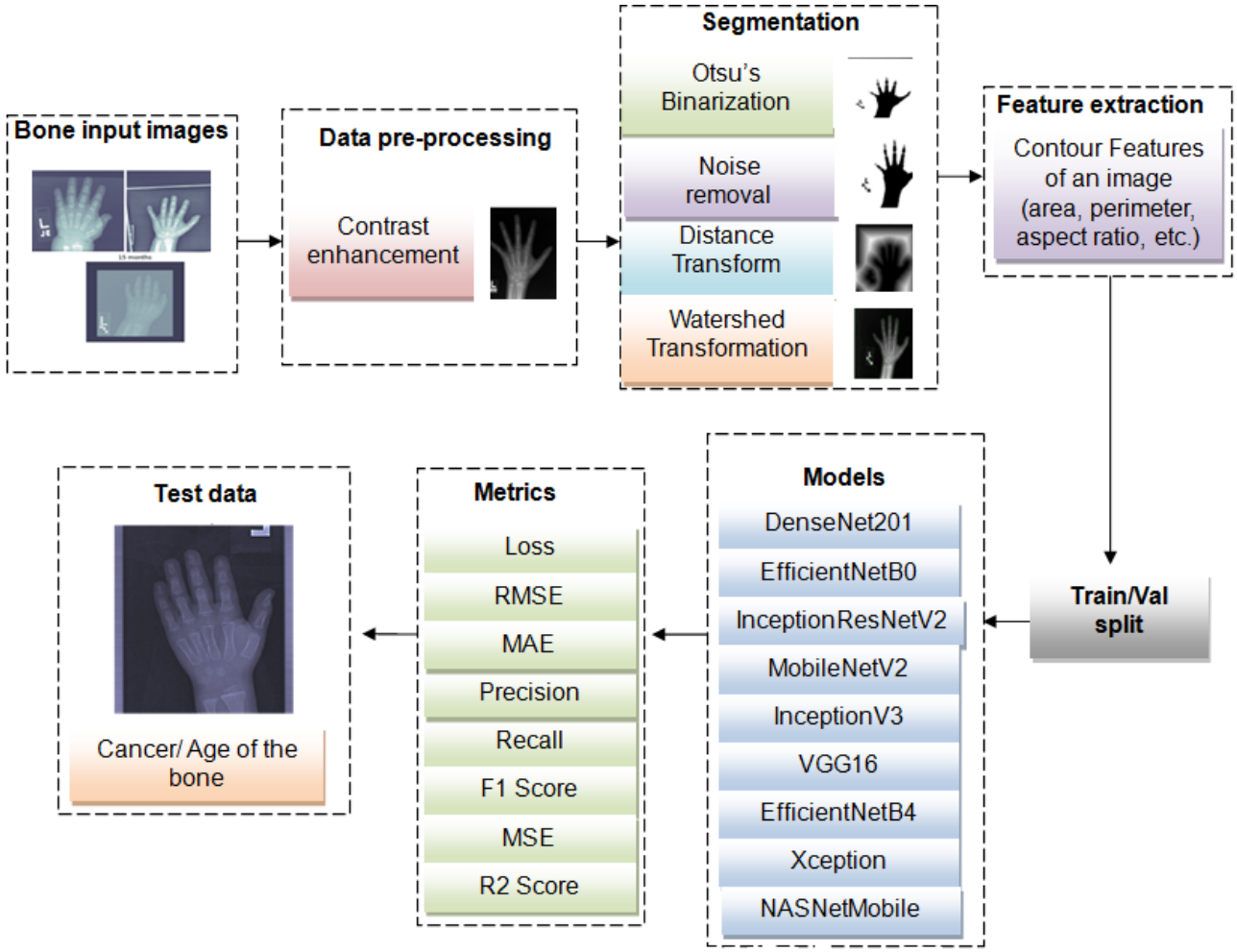

Fig 1. Proposed system to detect and predict bone cancer and age of the bone

dataset also includes detailed expert annotations in the form of bounding boxes, lines, polygons, and tags, derived from the X-ray images and corresponding medical reports. These annotations facilitate training of machine learning models for tasks such as fracture detection, classification, and localization [13]. Figure 2 illustrates sample images from the dataset.

2.2 Data pre-processing

Later, in the subsequent stage of image preprocessing, one widely used technique for the purpose of enhancing the contrast of images is histogram equalization, as shown in Figure 3. By redistributing the intensity values across the entire range, histogram equalization aims to create a more balanced distribution of pixel intensities. This adjustment effectively stretches the pixel value range, enhancing the differences between light and dark regions [14]. The application of contrast enhancement techniques is particularly beneficial in various fields, such as medical imaging, where it can aid in the identification of subtle abnormalities.

Let's denote the input grayscale image as $I(x, y)$ where x and y represent the spatial coordinates. The cumulative distribution function (CDF) of the image intensity values is given by (i):

$$CDF\ (i) = \sum_{j=0}^{i} H(j) \tag{i}$$

Where i represents the index, H(j) is the histogram of the image, representing the frequency of occurrence of intensity level j. The histogram equalization transformation function, T(i), can be defined as (ii):

$$T(i) = \frac{CDF(i) - CDF_{\min}}{MxN - 1}\ x\ (L-1) \tag{ii}$$

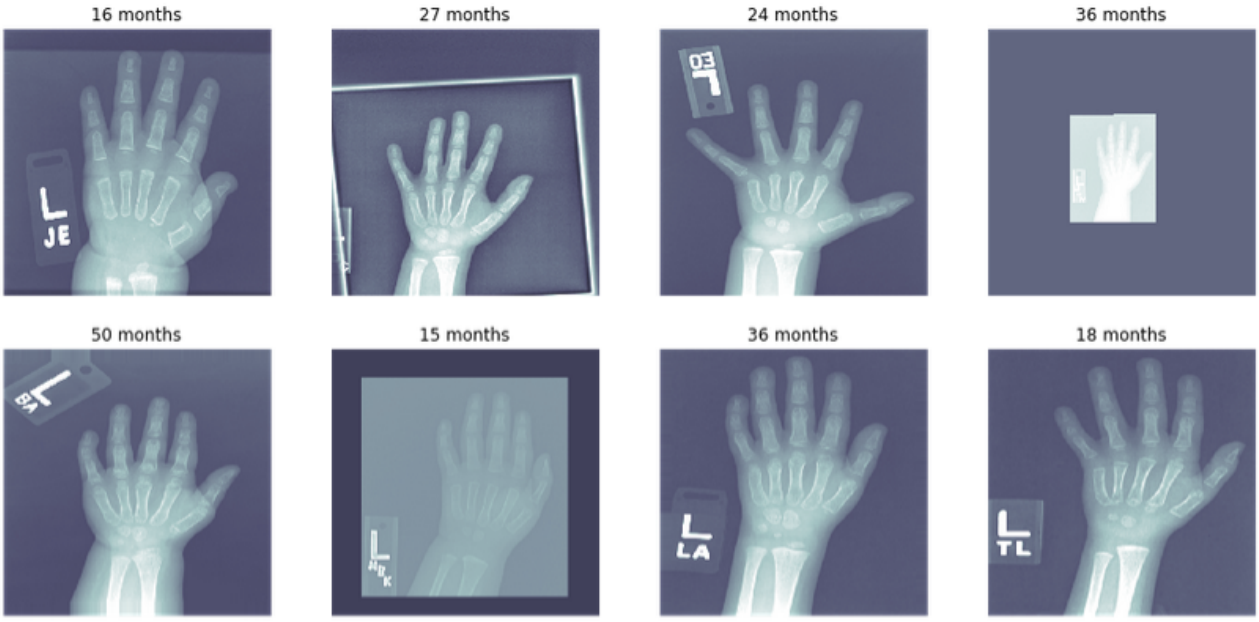

Fig 2. Original sample images of bone

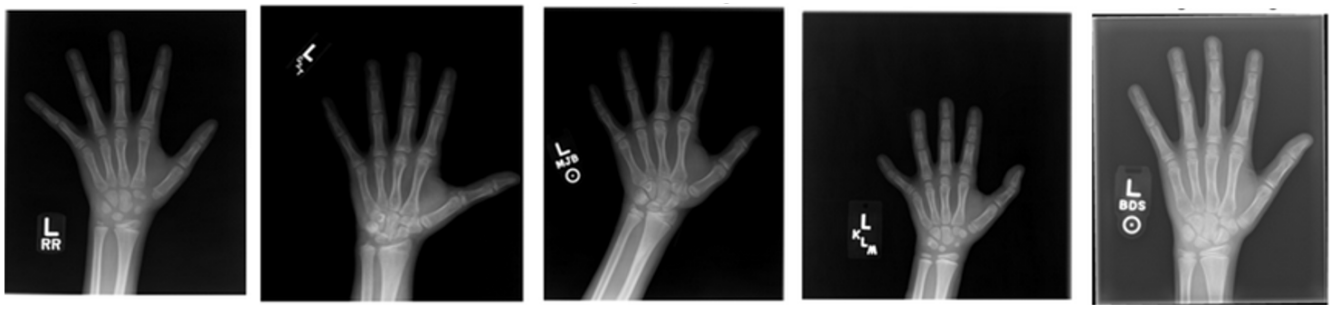

Fig 3. Quality enhanced of bone images

Here, i refer to the intensity level and L is the number of possible intensity levels, CDFmin is the min non-zero value of the CDF and M, N are the image dimensions. Finally, the histogram-equalized image, $I_{eq}(x,y)$is obtained by applying the transformation to each pixel (iii):

$$I_{eq}(x,y) = T(I(x,y)) \tag{iii}$$

This transformation effectively spreads the pixel intensity values across the entire range, leading to an enhanced contrast in the resulting image.

2.3 Segmentation

Otsu's Binarization, noise removal, Distance Transform, and Watershed Segmentation are interconnected stages in a comprehensive image segmentation process, as shown in Figure 4. Otsu's Binarization plays a pivotal role in converting a grayscale image into a binary format, separating objects from the background[15]. The algorithm maximizes the inter-class variance between background and foreground pixel intensities. The optimal threshold (T) is calculated using the following equation (iv):

$$T = argmax_t \left\{ \omega_1(t).\omega_2(t).[\mu_1(t) - \mu_2(t)]^2 \right\} \tag{iv}$$

Here, t refers to the potential threshold value, $\omega_1(t).\omega_2(t)$ are the probabilities of the two classes separated by the threshold, $\mu_1(t) - \mu_2(t)$ are the mean intensities of the two classes. The binary image is then subjected to noise removal techniques, ensuring the elimination of unwanted artifacts that could adversely impact subsequent segmentation steps[16]. Following noise removal, the Distance Transform is applied to calculate the distances of pixels to the nearest object boundaries[17]. There is a common formula for calculating distance transformation based on the Euclidean metric expressed by equation (v):

$$D(x,y) = \sqrt{\sum_{i=-k}^{k} \sum_{j=-k}^{k} I(x+i,\ y+j)x\ I(x+i,\ y+j)} \tag{v}$$

In this, D(x,y) is the distance value at pixel coordinates (x,y), I(x+i, y+j) refers to the intensity value of the pixel having the coordinates (x+i, y+j) in the original image, and k is the neighbourhood size used to define the local region around each pixel where distance is calculated. This distance map serves as a critical input for Watershed Segmentation, where the image is treated as a topographic surface. The watershed lines are defined based on the gradient of the distance map, and regions enclosed by these lines correspond to segmented objects[18]. Here is a simplified representation of the watershed transform equation in the context of grayscale images (vi):

$$W = ImposedMarkers \bigoplus RegionalMinima(GradientImage) \tag{vi}$$

In this case, W is the result of the watershed transform, ImposedMarkers are the positions of the markers, the Gradient Image is the gradient of the original grayscale image, and $\bigoplus$ is the morphologic reconstruction, which represents the basic operation in the fruition of the watershed transform.

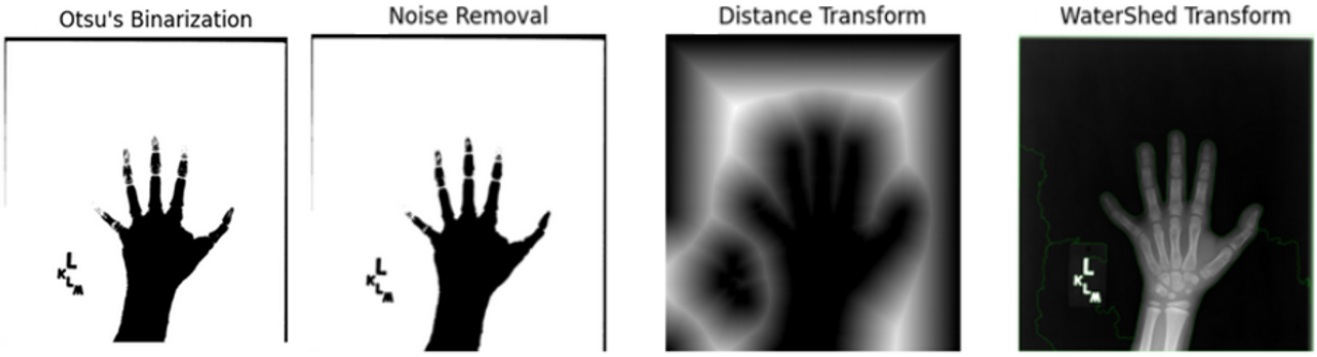

Fig 4. Image segmentation techniques applied on the bone X-ray images

2.4 Feature extraction

In the domain of image processing and computer vision, it is important to accurately characterize the shapes as well as the contours of various images as it helps to analyze and recognize the objects accurately[19]. Various set of parameters have been computed which quantify the different geometrical attributes such as area, perimeter, height, width, extent, aspect ratio etc. All these parameters are responsible for providing the unique in detail insights of the spatial characteristics of the contour of an image. Apart from this, mean color of the contour provides the deep analysis of both geometrical as well as pixel intensity aspects. Moreover, there are also few extreme points of the contour such as rightmost, leftmost, bottommost, and topmost points which have been calculated to offer important as well as required spatial information about the boundary of the desired shape[20].

Later, the data is split using a stratified approach to ensure that each subset maintains the same distribution of classes as the overall dataset. 70% of the data is given to training, 15% to validation, and the remaining 15% to testing. Such a division allows the model to gain experience on large amounts of data, and at the same time makes it possible to validate the learned parameters on the training samples. Also, a different and unseen test portion is provided to assess their accuracy objectively. The stratification step guarantees that both training and validation sets are drawn from the same class distribution as the overall class making it easier for the model to generalize to new data effectively.

2.5 Applied classifiers

In this segment, various deep learning classifiers employed for the estimation of bone disorders and bone age have been systematically discussed. These include DenseNet201, EfficientNetB0, EfficientNetB4, MobileNetV2, NASNetMobile, InceptionResNetV2, VGG16, Xception, and InceptionV3. Each of these models brings unique strengths to the task: DenseNet201 promotes feature reuse through dense connectivity, making it highly efficient in learning complex patterns; EfficientNetB0 and B4 scale depth, width, and resolution systematically to achieve better performance with fewer parameters; MobileNetV2 is optimized for mobile and embedded applications with its inverted residuals and linear bottlenecks; NASNetMobile leverages neural architecture search for optimal structure discovery; InceptionResNetV2 combines Inception modules with residual connections for faster convergence; VGG16 is a deep but simple model using uniform 3×3 convolutional layers; Xception replaces standard convolutions with depth wise separable convolutions for efficiency; and InceptionV3 uses asymmetric and factorized convolutions for reduced computational cost without compromising accuracy [17,21–25].

In all models, the input images are first processed through a series of convolutional layers that act as powerful feature extractors, transforming them into high-level abstract representations. Given the complexity and depth of these models, they can capture intricate structural patterns necessary for reliable diagnosis. To facilitate stable and accelerated training, Batch Normalization layers are applied to normalize intermediate activations, reducing internal covariate shift. Global Average Pooling is used to convert feature maps into compact vectors, effectively reducing dimensionality and preventing overfitting. During training, Dropout layers are incorporated to randomly deactivate neurons, forcing the network to learn redundant representations and thus generalize better. The final Dense layer maps the learned features to class probabilities or regression values depending on the task.

Among all the models, the layered architecture of the few models is shown. Table 1 provides a layered architecture of DenseNet201 network. Batch_Normalization_105 incorporates four trainable parameters and processes input data with dimensions (None, 512, 512, 1) followed by the extraction of features by DenseNet201 layer which results in an output shape of (None, 16, 16, 1920) with a count of 18,315,712. Later, the output is normalized by the Batch_Normalization_106 and spatial dimension is reduced by Global_average_pooling2d layer with 7680 and 0 as its trainable parameters respectively. Later, a 1D output is generated by having an output layer (None, 1) by Dense_5 layer with 1,921 trainable parameters. In total, the model comprises 18,325,317 parameters, with 18,092,419 being trainable and 232,898 non-trainable. As far as bone disease and bone age detection are concerned, this network has an ability in capturing the complex patterns to help in accurately identifying bone abnormalities and the estimation of bone age.

Table 1. Layered architecture of DenseNet201

Layer	Output Shape	**Param#**
$Batch\_Normalization\_105$	$(None, 512, 512, 1)$	4
$DenseNet201$	$(None, 16, 16, 1920)$	18315712
$Batch\_Normalization\_106$	$(None, 16, 16, 1920)$	7680
$Global\_average\_pooling2d$	$(None, 1920)$	0
$Dropout\_5$	(None, 1920)	0
$Dense\_5$	(None, 1)	1921
Total Params	18325317	
Trainable Params	18092419	
Non-trainable Params	232898	

Likewise, InceptionResNetV2 is a fusion of Inception and ResNet architectures which is characterized by its intricate and efficient feature extraction mechanisms to excel in discerning complex patterns. Table 2 outlines the architecture and key parameters of a neural network model, likely based on InceptionResNetV2, for a specific application, potentially related to bone disease or bone age detection. The model begins with a Batch Normalization layer applied to input images of shape (512, 512, 1). Subsequently, the InceptionResNetV2 architecture processes the input, that results in an output shape of (14, 14, 1536) with a total of 54,336,160 parameters. Following this, another Batch Normalization layer is employed, maintaining the output shape. The Global Average Pooling 2D layer condenses the spatial dimensions, yielding an output shape of (1536). A Dropout layer is then applied to prevent overfitting, with no parameters. Finally, a Dense layer outputs a single value, potentially indicative of the model's prediction for the given task. The model, in total, has approximately 54.3 million parameters, with 54.3 million being trainable and 63,618 being non-trainable.

Table 2. Layered architecture of InceptionResNetV2

Layer	Output Shape	Param#
$Batch\_Normalization\_107$	$(None, 512, 512, 1)$	4
$InceptionResNetV2$	$(None, 14, 14, 1536)$	54336160
$Batch\_Normalization\_108$	$(None, 14, 14, 1536)$	6144
$Global\_average\_pooling2d$	$(None, 1536)$	0
$Dropout\_6$	$(None, 1536)$	0
Dense_6	(None, 1)	1537
Total Params	54343845	
Trainable Params	54280227	
Non-trainable Params	63618	

At last, Table 3 presents the structure and essential characteristics of a EfficientNetB4 model which is designed for a particular purpose, potentially associated with bone disease or bone age detection. The model commences with a Batch Normalization layer that is applied to input photos with dimensions of (512, 512, 1). Afterwards, the EfficientNetB4 architecture analyzes the input, producing an output shape of (16, 16, 2048) and a total of 28,512,656 parameters. Subsequently, a subsequent Batch Normalization layer is implemented, preserving the output shape. The Global Average Pooling 2D layer reduces the spatial dimensions and Dropout layer mitigates the overfitting while as Dense layer produces the output. The model has 28.5 million parameters in which 28.3 million are trainable and 176,834 are non-trainable.

Table 3. Layered architecture of EfficientNetB4

Layer	Output Shape	Param
$Batch\_Normalization\_115$	$(None, 512, 512, 1)$	4
$EfficientNetB4$	$(None, 16, 16, 2048)$	28512656
$Batch\_Normalization\_116$	$(None, 16, 16, 2048)$	8192
$Global\_average\_pooling2d$	$(None, 2048)$	0
$Dropout\_10$	(None, 2048)	0
$Dense\_10$	(None, 1)	2049
$Total\ Params$	28522901	
$Trainable\ Params$	28346067	
$Non-trainable\ Params$	176834	

3 Result and Discussion

In this section, the performance of the applied models after being trained using bone cancer detection dataset, have been examined based on multiple metrics such as precision, root mean square error, recall, loss, F1 score, mean absolute error, and R2 score.

Initially the models are evaluated for loss and RMSE values during both training and validation phase as shown in Table 4. It can be found that EfficientNetB0 and DenseNet201 computed the lowest loss (0.0327 and 0.0268; 0.0325 and 0.0280 respectively) as well as RMSE values (0.180 and 0.163; 0.180 and 0.167 respectively) of in both the phases which showcases that these models exhibit robust generalization capabilities along with the effectively capturing of relevant features for accurate predictions. InceptionResNetV2 and NASNetMobile also present competitive results, with balanced Loss values for both training and validation datasets. InceptionResNetV2 achieved 0.0475 and 0.0349, and NasNetMobile recorded 0.0381 and 0.0550 for training and validation, respectively. These models demonstrate a good compromise between complexity and generalization. However, MobileNetV2 and EfficientNetB4 exhibit signs of potential overfitting, as evidenced by significant disparities between training and validation performance. MobileNetV2 showed a notable increase in Loss from 0.0441 during training to 0.1093 during validation, while EfficientNetB4 exhibited a drastic rise from 0.0536 to 0.6727 which was also the highest validation loss value among the other models. These models may struggle with adapting to unseen data which suggests a need for regularization techniques or model adjustments. VGG16 and Xception also display challenges in generalization, with considerable gaps between training and validation losses. The loss values of VGG16 were 0.0549 and 0.1649, and in case of Xception, the values

Table 4. Evaluative parameters of the applied models

Models	Training		Validation	
	Loss	RMSE	Loss	RMSE
DenseNet201	0.0325	0.180278	0.0280	0.167332
InceptionResNetV2	0.0475	0.217945	0.0349	0.186815
NasNetMobile	0.0381	0.195192	0.0550	0.234521
EfficientNetB0	0.0327	0.180831	0.0268	0.163707
MobileNetV2	0.0441	0.21	0.1093	0.330606
EfficientNetB4	0.0536	0.231517	0.6727	0.820183
VGG16	0.0549	0.234307	0.1649	0.406079
Xception	0.0649	0.254755	0.0597	0.244336
InceptionV3	0.0934	0.305614	0.0649	0.254755

were 0.0649 and 0.0597 for training and validation, respectively. InceptionV3, while providing satisfactory performance, has a higher training loss value as compared to the top-performing models. Its Loss values were 0.0934 for training and 0.0649 for validation. In addition to this, the models which also obtained the highest RMSE values during both training and validation phases are InceptionV3 and EfficientNetB4 with 0.305614 and 0.820183 respectively.

Figure 5 presents the training and validation loss curves of various deep learning models across epochs. In the case of DenseNet201 and EfficientNetB0, the training loss steadily decreases, indicating that the models are learning from the training data. However, their validation loss curves initially follow this trend but later show sharp increases, suggesting overfitting where the models perform well on training data but fail to generalize to new, unseen data. On the other hand, InceptionResNetV2 shows both training and validation loss decreasing and stabilizing together, indicating better generalization and consistent performance across both datasets. This suggests that while some models struggle with overfitting, others like InceptionResNetV2 achieve a more balanced learning behavior.

Additionally, for a thorough analysis, we have also extended our considerations beyond training times to encompass task-specific metrics such as precision, recall, and F1 score for the whole as well as both bone cancer and non-cancerous dataset.

Figure 6 illustrates and compares the precision, recall, and F1 scores of various deep learning models on the complete dataset. From the interpretation, it is evident that DenseNet201 is highly precise (98.5%), meaning it makes very accurate positive predictions, but its lower recall and F1 score indicate it may miss some true cases. InceptionResNetV2 emerges as one of the best-performing models, with a strong balance across all metrics, especially its high F1 score (98.21%), suggesting excellent generalization and minimal false negatives. NASNetMobile also shows consistent and balanced performance, with all metrics close in value, making it a reliable model. EfficientNetB0 and EfficientNetB4 demonstrate strong precision but slightly lower recall, affecting their F1 scores. MobileNetV2 maintains uniform performance across precision, recall, and F1, indicating stability. Models like VGG16 and InceptionV3 show good recall but somewhat weaker precision and F1 scores, implying limitations in handling more complex cases. Xception has high recall but low precision, resulting in a lower F1 score—indicating a tendency to over-predict positive cases and generate more false positives.

From Table 5, the evaluation of bone cancer detection models reveals varying strengths across architectures. DenseNet201 demonstrated strong precision, indicating high reliability in positive predictions, while maintaining a reasonable recall. InceptionResNetV2 delivered a well-balanced performance across all metrics, showcasing its robustness. NasNetMobile and EfficientNetB0 showed competitive results, with NasNetMobile slightly edging out in recall and overall F1 score. MobileNetV2 emerged as a top performer in recall, contributing to a high F1 score, suggesting its effectiveness in identifying true positive cases. Although EfficientNetB4 excelled in precision, its comparatively lower recall and F1 score indicate possible limitations in generalization. VGG16 offered balanced performance, whereas Xception stood out for its high recall. Notably, InceptionV3 achieved the highest recall and an impressive F1 score, highlighting its superior sensitivity and overall effectiveness in detecting bone cancer cases.

In the assessment of deep learning models for detecting non-cancerous bone cases, DenseNet201 achieved the highest precision, indicating strong performance in minimizing false positives, though its recall suggests some missed true positives. In contrast, InceptionResNetV2 delivered an outstanding balance between precision and recall, as reflected in its near-perfect F1 score. Both NASNetMobile and EfficientNetB0 maintained strong overall performance, combining high precision with reasonable recall. EfficientNetB4 also performed well, with a high F1 score indicating a consistent balance of sensitivity and specificity. Similarly, MobileNetV2 and VGG16 demonstrated competitive results, effectively balancing both evaluation metrics.

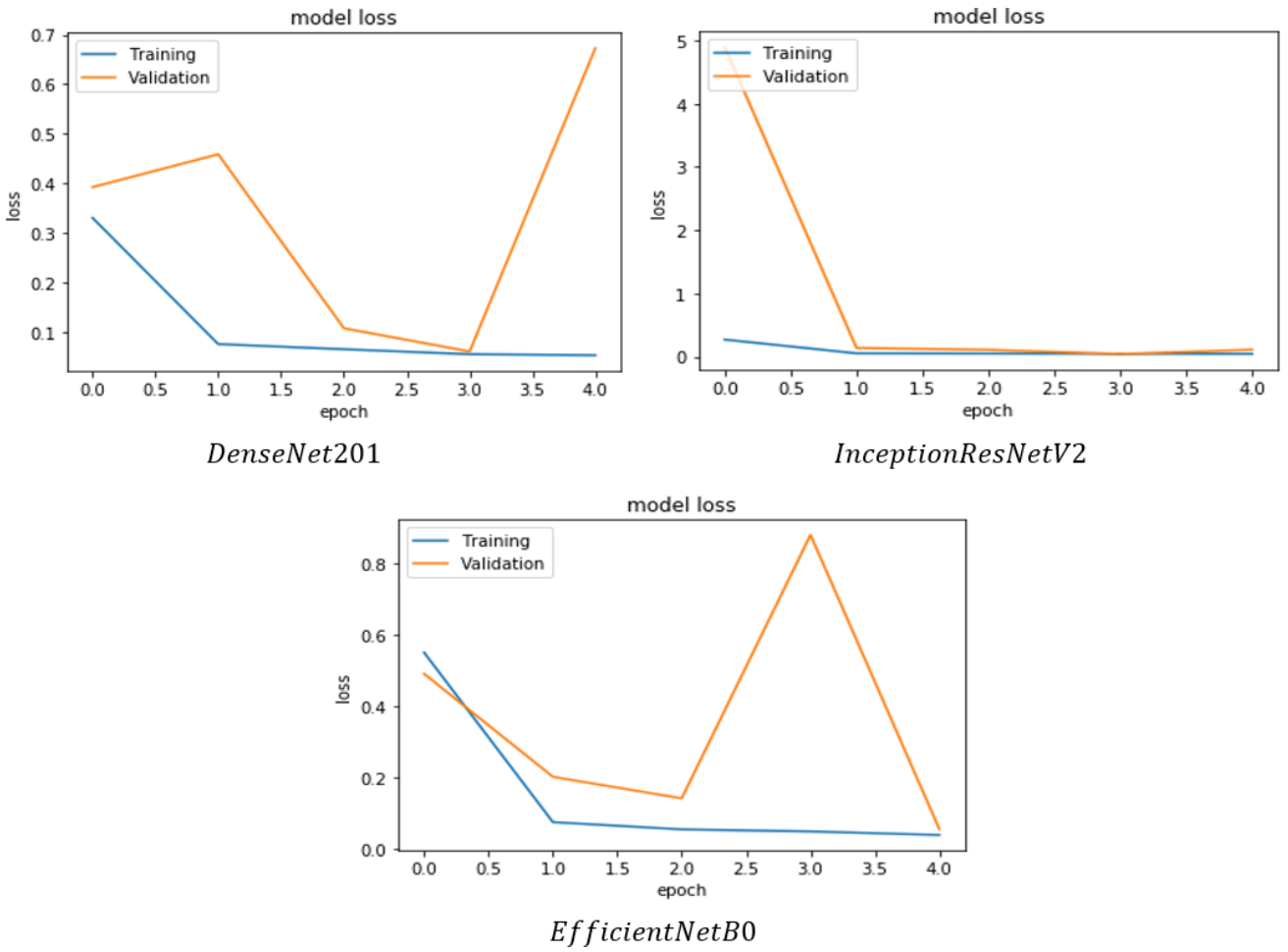

Fig 5. Learning curves of the applied models

Table 5. Examination of the models for cancerous and non-cancerous bone data

Models	Bone Cancer			Non-Cancerous		
	Precision	Recall	F1Score	Precision	Recall	F1Score
DenseNet201	98.00	90.00	93.46	99.00	93.47	90.58
InceptionResNetV2	94.64	93.48	96.56	91.59	95.55	99.86
NasNetMobile	93.46	94.34	94.57	98.64	93.48	93.46
EfficientNetB0	95.33	93.46	93.48	97.66	92.22	92.58
MobileNetV2	91.66	93.59	97.87	94.54	93.24	92.16
EfficientNetB4	97.56	92.15	90.34	99.36	93.14	95.56
VGG16	89.99	94.53	91.49	89.46	92.56	92.56
Xception	85.42	93.57	82.62	79.48	94.58	86.49
InceptionV3	88.69	99.36	98.49	89.49	91.56	82.56

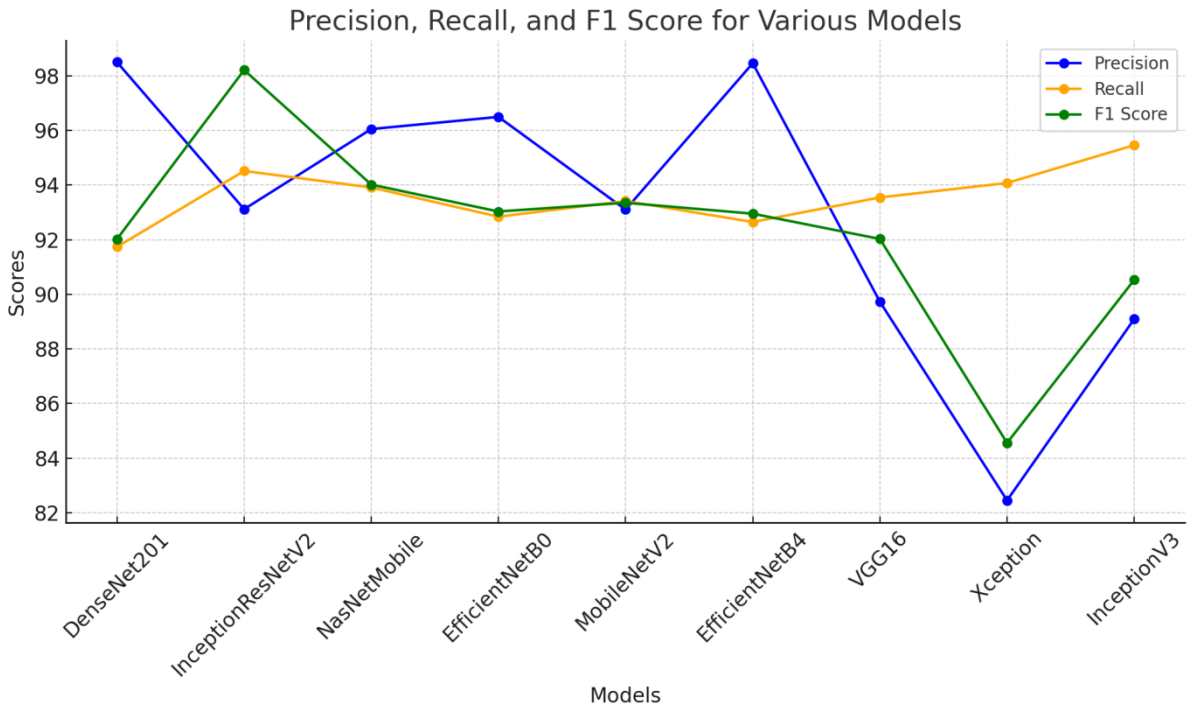

Fig 6. Performance metrics of the applied deep learning models

Xception, while achieving a high recall, suffered from lower precision, leading to a reduced F1 score. Meanwhile, InceptionV3 showed moderate performance across all metrics, lacking the distinct strengths observed in other models. Overall, the results highlight varying trade-offs between precision and recall, with some models offering more balanced and reliable predictions than others.

In addition to this, the execution or performance of the applied models for the validation dataset of bone age have been also examined on the basis of other parameters such as Mean squared error, Mean absolute error, and R2 score, as shown in **Figures 7, 8, and 9**.

In evaluating the performance of various deep learning models for image classification, **Mean Squared Error (MSE)** is used as a metric. The results indicate distinct performances among the considered models. NASNetMobile demonstrates the lowest MSE at 153.03 which suggests the superior predictive accuracy. EfficientNetB0 and MobileNetV2 follow closely with MSE values of 168.49 and 176.31, respectively, thereby showcase their competitive performance. However, DenseNet201 and EfficientNetB4 exhibit relatively higher MSE values at 267.09 and 267.59, which indicate a comparatively less accurate predictive capability. InceptionV3, VGG16, and Xception fall within the intermediate range with MSE values of 244.20, 246.88, and 256.44, respectively. These findings collectively suggest that NASNetMobile, EfficientNetB0, and MobileNetV2 are promising models for image classification tasks based on their lower MSE values, while further investigation and optimization may be warranted for models with higher MSE, such as DenseNet201 and EfficientNetB4.

On evaluating the models based on **Mean Absolute Error (MAE),** DenseNet201 obtained the lowest MAE of 13.56 which indicates its accurate prediction capability. Xception, InceptionV3, and MobileNetV2 exhibit great results, with MAE values of 18.66, 18.16, and 18.33, respectively. VGG16 closely follows with an MAE of 14.59 which suggested a strong performance as well. NASNetMobile and InceptionResNetV2 fall within the mid-range that demonstrates the MAE values of 15.33 and 16.59. However, models such as EfficientNetB4 and EfficientNetB0 show relatively higher MAE values at 19.49 and 26.59, respectively which indicate their potential areas for improvement in their predictive accuracy. Overall, DenseNet201 and VGG16 showcase strong performance based on their low MAE values, while further optimization may be explored for models with higher MAE, such as EfficientNetB0 and EfficientNetB4.

Evaluation based on the R^2 score revealed that MobileNetV2 achieved the highest score of 96.59, indicating its superior capability to explain variance in bone age prediction. EfficientNetB4 (95.49), Xception (94.56), and DenseNet201 (94.59) also demonstrated strong predictive performance. In contrast, InceptionV3 recorded the lowest R^2 score of 86.29, suggesting

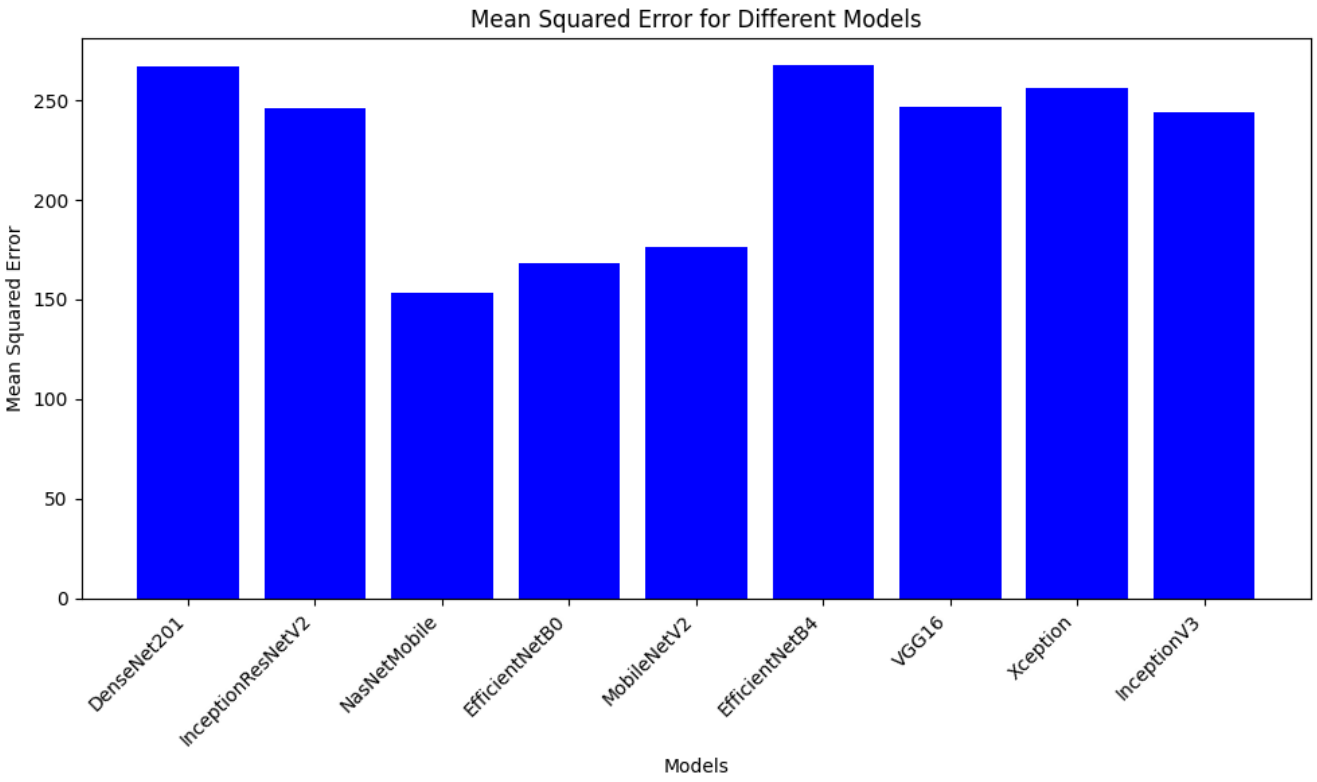

Fig 7. MSE evaluation of the applied CNN models

limited ability to capture data variability. Models like VGG16 (93.26) and EfficientNetB0 (92.59) performed competitively, while InceptionResNetV2 (91.26) and NASNetMobile (90.49) also yielded respectable results. Overall, MobileNetV2, DenseNet201, EfficientNetB4, and Xception emerged as top performers for image-based bone age prediction tasks.

Table 6 summarizes key recent studies in bone cancer detection and bone age estimation, highlighting their datasets, techniques, and performance metrics. The proposed study demonstrates clear advantages in accuracy, generalizability, and dual-task capability over prior approaches.

4 Conclusion

The paper introduces a robust deep learning system capable of both bone cancer detection and bone age estimation within a unified framework. By integrating advanced preprocessing methods such as histogram equalization, Otsu thresholding, distance transformation, watershed segmentation, and contour-based feature extraction, the model enhances input quality and emphasizes critical image regions. This preprocessing pipeline enables the network to capture fine structural details, leading to superior performance across multiple evaluation metrics and demonstrating clear improvements over prior approaches.

The scope of this study is confined to automated analysis of pediatric bone X-rays for the dual tasks of bone cancer detection and skeletal age estimation.

Quantitatively, the proposed system outperforms prior studies in both bone cancer detection and age estimation. DenseNet201 achieved a 93.46% F1 score for cancer detection and the lowest MAE (13.56) for age prediction, while EfficientNetB0 yielded the best RMSE (0.1637) in age estimation. InceptionResNetV2 showed exceptional robustness with a 98.21% F1 score and consistent performance. These results, based on a large dataset of 12,611 pediatric bone X-rays and validated on unseen clinical images, significantly surpass earlier models like Xinyang et al. (2024), which used traditional approaches and smaller datasets. High R^2 values (>94.5%) further confirm the models' reliability for clinical applications.

This paper introduces a dual-purpose AI system capable of both disease classification and age estimation from a single image. By combining deep learning with image enhancement and region-specific preprocessing, it improves diagnostic accuracy while minimizing manual effort. However, limitations include overfitting in deeper models like EfficientNetB4 (training loss: 0.0536

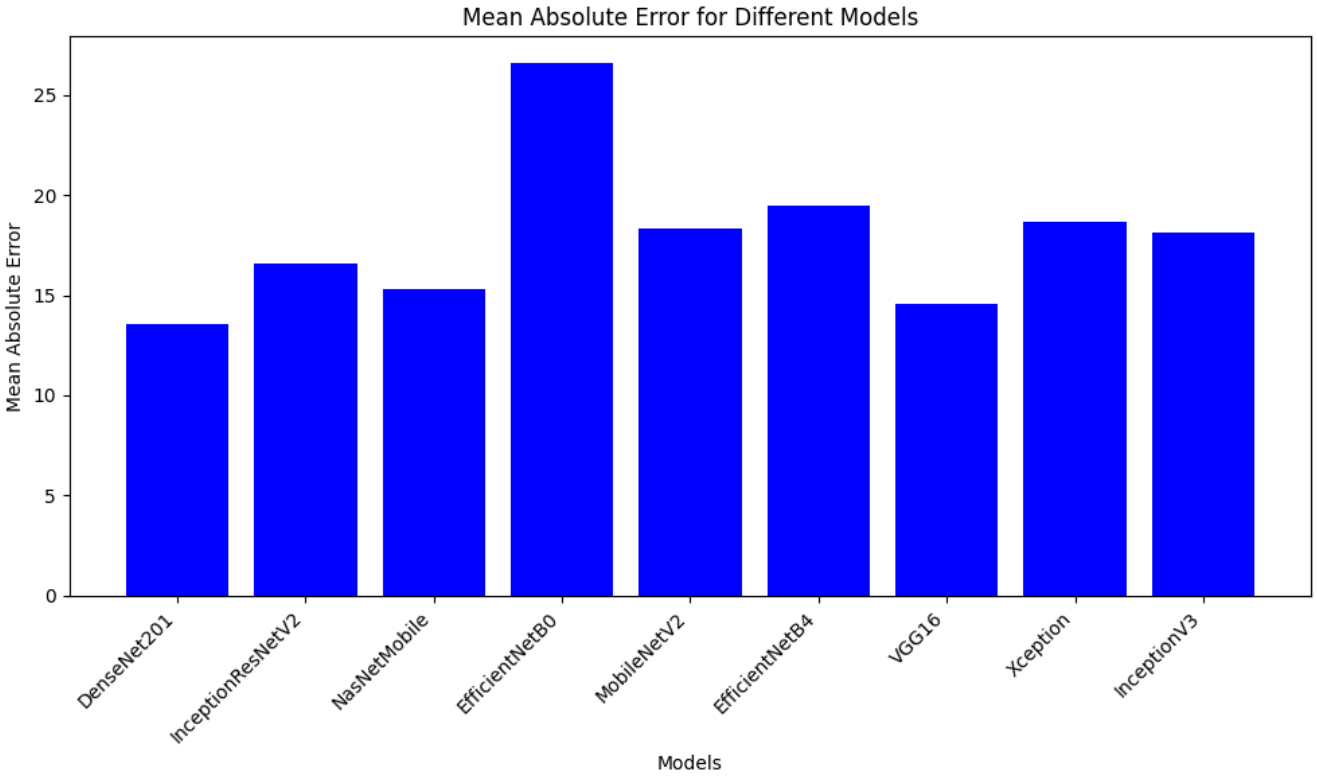

Fig 8. MAE evaluation of the applied CNN models

Table 6. Comparison of the proposed study with recent related works

Study	Dataset	Technique	Results	Limitations	Improvement
Xinyang et al. (2024) [4]	Few hundred MRI images	Radiomics + ML	Accuracy: 82.8%	Focused on prostate metastasis; no deep features	F1: 93.46% (DenseNet201); large multi-class dataset (12,611 images)
Choi et al. (2024) [11]	5,000 pediatric X-rays	Contrast-enhanced CNNs	RMSE: 0.42	Poor generalization on unseen data	RMSE: 0.1637 (EfficientNetB0); >60% error reduction
Gawade et al. (2023) [6]	1,000 single-cancer images	CNN + SVM	Accuracy: 87%	No segmentation; weak cross-class generalization	F1: 98.21% (InceptionResNetV2); includes advanced segmentation
Ramasamy et al. (2023) [10]	3,000 histopathology images	Hybrid CNN	Accuracy: 99.45%	Single cancer type; no regression	Adds bone age prediction; 9-model validation with high accuracy
Sahin (2023) [7]	2,000 X-rays	Classical ML + Image processing	High precision	No deep learning; poor unseen data performance	Precision: 98.5%, Recall: 90%; deep CNNs + contour segmentation
Anand et al. (2022) [8]	2,500 images	DC-ELM	Accuracy: 91%	Handcrafted features; limited flexibility	Fully automated; superior accuracy via deep transfer learning

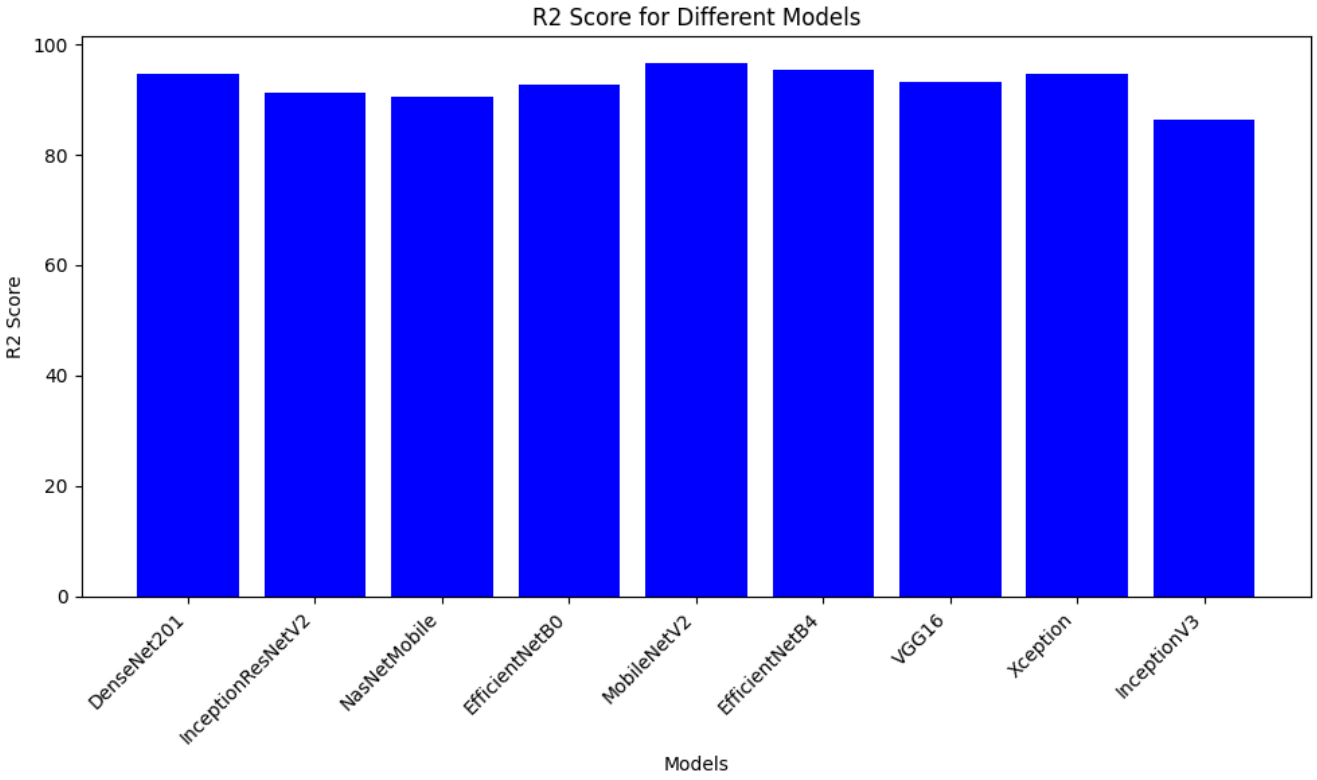

Fig 9. R2 score evaluation of the applied CNN models

vs. validation loss: 0.6727), class imbalance affecting cancer detection sensitivity, lack of clinical validation, and dataset sourcing from a single institution, which may impact generalizability.

The future direction of this research emphasizes expanding the dataset across hospitals and imaging modalities (MRI, CT), using augmentation techniques like GANs or SMOTE to mitigate overfitting and imbalance, and applying explainable AI tools (e.g., Grad-CAM, SHAP) for transparency. Collaboration with clinicians is crucial for validation and ethical deployment. In doing so, this work contributes a novel dual-task framework and introduces clinically significant improvements over previous studies, thereby extending the boundaries of current medical image analysis literature.

References

1) Cavallo F, Mohn A, Chiarelli F, Giannini C. Evaluation of bone age in children: a mini-review. *Frontiers in Pediatrics*. 2021;9:580314. https://doi.org/10.3389/fped.2021.580314.
2) Bone cancer - Symptoms and causes - Mayo Clinic. 2023.
3) Professional, C. C. M., Bone Cancer. .
4) Xinyang S, Shuang Z, et al. A machine learning radiomics model based on bpMRI to predict bone metastasis in newly diagnosed prostate cancer patients. *Magnetic Resonance Imaging*. 2024;107:15–23. https://doi.org/10.1016/j.mri.2023.12.009.
5) Jeong Y, Jeong C, et al. Development of AI-based diagnostic algorithm for nasal bone fracture using deep learning. *Journal of Craniofacial Surgery*. 2024;35(1):29–32. https://doi.org/10.1097/SCS.0000000000009856.
6) Gawade S, Bhansali A, Patil K, Shaikh D. Application of the convolutional neural networks and supervised deep-learning methods for osteosarcoma bone cancer detection. *Healthcare analytics*. 2023;3:100153. https://doi.org/10.1016/j.health.2023.100153.
7) Sahin ME. Image processing and machine learning-based bone fracture detection and classification using X-ray images. *International Journal of Imaging Systems and Technology*. 2023;33(3):853–865. http://dx.doi.org/10.1002/ima.22849.
8) Anand D, Arulselvi G, Balaji GN. A deep convolutional extreme machine learning classification method to detect bone cancer from histopathological images. *Journal of Optoelectronics Laser*. 2022;41(7):456–468. Available from: https://ijisae.org/index.php/IJISAE/article/view/2194.
9) Suganeshwari G, Balakumar R, et al. DTBV: A deep transfer-based bone cancer diagnosis system using VGG16 feature extraction. *Diagnostics*. 2023;13(4):757. https://doi.org/10.3390/diagnostics13040757.
10) Ramasamy MD, Dhanaraj RK, et al. An improved deep convolutionary neural network for bone marrow cancer detection using image processing. *Informatics in Medicine Unlocked*. 2023;38:101233. https://doi.org/10.1016/j.imu.2023.101233.

11) Choi DH, Ahn SH, Lee R. An accurate pediatric bone age prediction model using deep learning and contrast conversion. *The Ewha Medical Journal*. 2024;47(2). https://doi.org/10.12771/emj.2024.e23.
12) Kassem MA, Naguib SM, et al. Explainable Transfer Learning-Based Deep Learning Model for Pelvis Fracture Detection. *International Journal of Intelligent Systems*. 2023;2023(1):3281998. https://doi.org/10.1155/2023/3281998.
13) Nagy E, Janisch M, et al. A pediatric wrist trauma x-ray dataset (grazpedwri-dx) for machine learning. *Scientific data*. 2022;9(1):222. https://doi.org/10.6084/m9.figshare.14825193.v2.
14) AL-Azzawi ZH, Mahdi WH, Ahmed SK, Yasin WS. Medical image enhancement using histogram equalization techniques;vol. 2591. 2023.
15) Yang P, Song W, Zhao X, Zheng R, Qingge L. An improved Otsu threshold segmentation algorithm. *International Journal of Computational Science and Engineering*. 2020;22(1):146–153. https://doi.org/10.1504/IJCSE.2020.107266.
16) Hooda M, Bachu S. Artificial intelligence technique for detecting bone irregularity using fastai. 2020. https://doi.org/10.46254/AN10.20200037.
17) Walid MA, Shill PC. A Transfer-Learning Based Unbiased Voting Bone Cancer Detection Framework from Histological Osteosarcoma Images. 2023.
18) Ma J, Chen J, Ng M, Huang R, Li Y, Li C, et al. Loss odyssey in medical image segmentation. *Medical image analysis*. 2021;71:102035. https://doi.org/10.1016/j.media.2021.102035.
19) Anusha N, Ramana MV. A Recent Survey On Bone Tumor Detection And Classification Using Deep Learning Techniques. *IOSR Journal of Computer Engineering (IOSR-JCE)*. 2025;27(1):28–32. https://doi.org/10.9790/0661-2701032832.
20) Aarthy R, Muthupriya V, Balaji GN. Detection of bone cancer based on a four-phase framework generative deep belief neural network in deep learning. *Alexandria Engineering Journal*. 2024;109:394–407. https://doi.org/10.1016/j.aej.2024.08.094.
21) Fan F, Liu H, Dai X, et al. Automated bone age assessment from knee joint by integrating deep learning and MRI-based radiomics. *Int J Legal Med*. 2024;138:927–938. https://doi.org/10.1007/s00414-023-03148-1.
22) Islam ME, Syed MA, et al. Pediatric Bone Age Prediction Using Deep Learning. 2023.
23) Barhoom AM, Al-Hiealy MR, Abu-Naser SS. Deep Learning-Xception Algorithm for upper bone abnormalities classification. *Journal of Theoretical and Applied Information Technology*. 2022;100(23):6986–6997.
24) Chen W, Ayoub M, et al. A fusion of VGG-16 and ViT models for improving bone tumor classification in computed tomography. *Journal of Bone Oncology*. 2023;43:100508. https://doi.org/10.1016/j.jbo.2023.100508.
25) Wang R, Zhao J, et al. AdvYOLO: Advanced YOLOv8 Application for Bone Pathology Localization and Classification in Wrist X-ray Images. 2024. http://dx.doi.org/10.21203/rs.3.rs-4051336/v1.