RESEARCH PAPER

# Automated Pediatric Bone Age Prediction Using Multitask Deep Learning Techniques

**Rumana Anjum[1], Dr. Madhu B K[2], Dr. Reshma Anjum[3]**
[1] Assistant Professor, Computer Science and Engineering, Vidya Vikas Institute of Engineering and Technology, Mysore, Karnataka. Email: rumanaanjum85@gmail.com
[2] Professor and Dean, Computer Science and Engineering, Vidya Vikas Institute of Engineering and Technology, Mysore, Karnataka. Email: professormadhubk@gmail.com
[3] Radiologist, Radiology, Mallige Diagnostic Centre, Mysore, Karnataka. Email: reshamid78@gmail.com

**Received:** 2nd Mar, 2026 | **Revised:** 14th Mar, 2026 | **Accepted:** 4th Apr, 2026 | **Available Online:** 20th Apr, 2026

## ABSTRACT

Bone age assessment is an important diagnostic tool in pediatric medicine to determine growth and developmental disorders. The proposed research project introduces a deep learning-driven, multitask, clinically reliable automated model to estimate the age of children's bones. The model combines both classification and regression into an EfficientNetB4-based framework to forecast developmental stages and accurately determine bone age. A structured preprocessing pipeline and a gender-based input representation are also introduced to enhance feature extraction. Also, there is a six-layer validation framework and a three-layer safety mechanism to guarantee the quality of input and quality predictions. Another aspect that is introduced in the system is the confidence estimation module that will be used in the process of clinical decision-making. Trained on the RSNA pediatric dataset, the model has a competitive performance with the ability to be interpreted and robust. The proposed model is clinical usability-oriented, safety-oriented, and deployment-ready, unlike the traditional and regression-only models. Overall, the research provides a solution that is scalable and interpretable and can bridge the gap between the artificial intelligence models and the reality of the field of pediatric healthcare.

**Keywords:** Bone Age Prediction, Multitask Learning, Deep Learning, EfficientNetB4, Pediatric Radiology, Medical Image Analysis, Clinical Decision Support.

**How to cite this article:** Anjum R, Madhu B K, Anjum R. Automated Pediatric Bone Age Prediction Using Multitask Deep Learning Techniques. Int J Drug Deliv Technol. 2026;16(34s):715-728. DOI: 10.25258/ijddt.16.34s.89

**Source of support:** Nil.

**Conflict of interest:** The authors declare no conflict of interest.

Age estimation of the bones is an important diagnostic tool in the medical care of children and provides valuable data about the skeletal maturity of a child and the overall developmental status(1). It is widely used in the diagnosis of growth diseases, endocrine diseases, developmental retardation, in the design of orthopaedic surgery, and in the evaluation of treatment results. The bone age, in contrast to the chronological age, which is all about the date of birth, is a measure of biological maturity and a more representative measure of physiological maturity(2). The left hand and wrist radiography is the most common clinical method of examination because several ossification centers vary in relation to age. The timely and accurate evaluation of bone age is crucial in the diagnosis of diseases like growth hormone deficiencies, precocious puberty, and constitutional growth delay, and allows medical intervention and better patient outcomes(3). However, the traditional methods of bone age determination, primarily the Greulich and Pyle (GPU) atlas method and the Tanner Whitehouse (TW) scoring process, have several limitations, which restrict their usefulness and applicability in practice today(4). The principle of the GP method consists of the visual comparison of the radiograph of an individual to the normal atlas pictures, which is highly subjective and depends on the experience of clinicians. Similarly, the TW method, more rigorous, involves a grading of individual bones in detail, is time-consuming, and prone to inter and intra-observer error. Both techniques need a lot of training and experience, and therefore, they are not consistent in interpretation, nor are they less reproducible with various practitioners and institutions. In addition, they cannot be used in large-scale screening or high-throughput clinical settings, as traditional methods are not automated and require manual work(5). To overcome these limitations, recent approaches have centered on deep learning-based approaches, which can be applied to automate feature extraction and prediction processes that can be assisted using convolutional neural networks (CNNs)(6). Although these models have proven to be more accurate and efficient in comparison to the traditional methods, they are not devoid of challenges. The developmental stages of skeletal growth are not accounted for in most present deep learning models, and instead, only age is regressed, which could lead to a lack of interpretability and clinical confidence. In addition, these models have low responses to variations in picture quality, position of the hand, or non-standard radiographs, resulting in inaccurate predictions(7). Another significant problem is the absence of inbuilt validation and safety controls because, without them, it is impossible to deploy it to clinical practice, and only the correct radiographs are to be analyzed. Moreover, most of the models do not have confidence estimation, and clinicians can hardly tell whether the predictions made by the models are reliable or not. Such systems are also hampered by their unpredictability, lack of generalization in various populations, and failure to fit into real clinical practices. This way, despite the advancement of artificial intelligence, there is still a need to have a purposeful, clinically sound, and explainable form that must overcome these limitations and bring about accuracy and strength, safety, and practicability in bone age estimation in paediatrics(8). Despite the tremendous advances in automated bone age estimation by deep learning techniques, several critical issues remain that prevent their application in the clinical community and their reliability(1). The current models are mostly concerned with single-task regression systems, which directly predict bone age without taking into account the systematic sequence of skeletal development stages. Such simplification usually leads to lower interpretability and discrepancy in forecasts, especially in varied pediatric groups(9). Furthermore, a majority of the existing approaches lack rigorous validation steps to ensure the quality of input images, proper anatomic positioning and applicability of radiographs, which are required to apply them safely in the clinical practice environment. The lack of built-in safety layers complicates the possibility of false forecasts being made in case of non-standard or low-quality images. Also, most models do not have the capability to estimate confidence, and clinicians can barely tell the reliability of automated predictions(10). The second weakness is that these systems have not been end-to-end integrated with end-user-friendly interfaces and deployment-ready architectures. These shortcomings lead to the need for a more comprehensive and clinically focused solution(11). The motivation for this research is therefore to develop a powerful, explainable, and implementable framework that not only improves prediction accuracy but also ensures the reliability, safety, and usefulness of Pediatric health care environments by integrating multitask learning, a validation pipeline, and confidence-aware prediction systems. This study aims to create an automated bone age prediction system in paediatrics based on a multitask deep learning framework where classification and regression are combined to achieve higher accuracy and interpretability(12). The research intends to use an effective CNN-based backbone to extract features, introduce developmental bucket strategies to simulate the skeletal growth stages, and introduce a validation pipeline to maintain the quality and safety of inputs(13). Also, the system incorporates the estimation of confidence to increase reliability and is modeled as an end-to-end deployable solution that can be used by clinicians.

## Data and Materials.

### RSNA Pediatric Bone Age Dataset Overview

The RSNA Pediatric Bone Age Challenge dataset is the dataset used in the study because it is a standard and commonly used benchmark dataset in automated bone age prediction activities that are automated. This dataset consists of a total of 12,611 hand radiographs, which are annotated with bone age values, which were obtained by expert radiologist consensus(14). The radiographs have largely focused on the left hand and left wrist as they provide several ossification centers that are highly significant in the determination of skeletal maturity(15). In addition to the age labels of the bones, some sex data are also available in the dataset that plays a significant role in skeletal development and is considered in the model to increase the quality of prediction. The dataset covers a wide pediatric age distribution, which allows the model to learn developmental trends since infancy to late adolescence. This is because of its high-quality and extensive annotations, clinical relevance, and the scale that it is very appropriate to train and evaluate the deep learning models targeted at automated bone age assessment(16).

### Data Distribution and Demographics

The dataset has a balanced distribution of demographic features, which guarantees the objective model training and testing. The dataset is skeletally mature in children aged between 1 month and 228 months (approximately 19 years), which encompasses all the important stages of skeletal development in children. The sample gender ratio is about equal (almost 50 percent of samples are male, and the other half are female), which is useful in minimizing gender bias and enabling the model to pick up sex-specific growth patterns. Furthermore, the distribution of data by stages of development is classified into age groups that are clinically significant, i.e., infancy, early childhood, mid-childhood, adolescence, and near maturity. These development buckets reveal the biological order of skeletal development and play a vital role in the multitask learning model of this study. The sampling distribution of these buckets ensures that such buckets give the model a large variety of various stages of growth to make predictions with more generalization and power.

### Preprocessing and Normalization of Data.

The raw radiographic images are processed in an end-to-end pipeline to transform them into an input format that is readable by a deep learning model. It begins with the decoding of images in the traditional format, such as PNG or JPG, and their conversion into grayscale to highlight structural data, and make computations easier. The pictures are then resized to a standard size of 300x300 pixels using the bilinear interpolation to ensure that the pictures are all the same size across the dataset. The values of pixels are scaled to the range [0,1], a factor that enhances the numerical stability in the training of the model. Moreover, Efficient Net-specific preprocessing is used, such as channel-wise normalization, and gray images are transformed into tri-channel images to align with the model input needs. A validation mechanism has also been included in the preprocessing pipeline to make sure that only high-quality radiographic images are processed. Moreover, the age values of bones are normalized with dataset-specific normalization parameters, such as the mean of age (127.394 months) and standard deviation of age (41.1211 months) to enable effective learning and quicker convergence during training.

### Dataset Partitioning Strategy

In order to guarantee a good training of the model and a fair estimation, the dataset is stratified with a stratified sampling strategy that equally represents the age and sex groups. In training, 90 percent of the dataset (11,349 images) is utilized and in validation, 10 percent (1,262 images) is utilized. This stratification makes sure that the two subsets are similar in terms of demographic and developmental attributes to avoid skewed learning or evaluation bias. The validation sample is sufficient to provide good performance assessment and this is especially so when we compare it to smaller benchmark test samples used in the past. This division approach is in favor of good generalization of models and appropriate projection of performance in the real world. The characteristics of the dataset are detailed and summarized in Table 2. which includes important information, including the number of total images, the range of ages, gender distribution, and the partition sizes. Also, Figure 2 shows the distribution and prediction behavior of the development at the various age stages, how the model has managed to capture the skeletal maturity patterns at the various stages.

| Characteristic | Value |
|---|---|
| Total Images | 12,611 hand radiographs |
| Age Range | 1-228 months (1 month to 19 years) |
| Gender Distribution | ~50% male, ~50% female |
| Training Set (90%) | 11,349 images |
| Validation Set (10%) | 1,262 images |
| Partitioning | Stratified sampling (age + sex balanced) |
| Normalization: AGE_MEAN | 127.3994 months |
| Normalization: AGE_STD | 41.1211 months |

**Table 2:** Dataset Characteristics

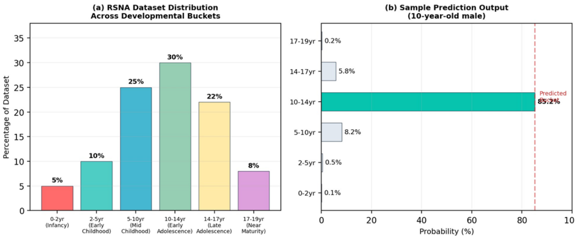

**Figure 2:** Developmental Bucket Distribution and Sample Prediction

### Preprocessing and Extraction of Features.

### Comparative Feature Extraction Model Analysis.

An extensive analysis of several deep learning models was done to find the best feature extraction model that suits pediatric bone age prediction. The most popular convolutional neural network architectures, such as VGG16, ResNet50, InceptionV3, MobileNetV2, and Efficient Net versions, were contrasted in terms of parameter size, computational complexity, input resolution, and feature extraction. VGG16, although having a good representational power, has about 138 million parameters, which is computationally intensive and is not applicable to scalability or real-time scenario.ResNet50 and InceptionV3 offer better performance based on residual connections and multi-scale feature extraction, respectively, but do not have effective scaling mechanisms. Lightweight models like MobileNetV2 minimize the amount of computation, although they might not resolve fine-grained skeletal details. In comparison with it, Efficient Net models have a scaling approach to the network, which balances the depth, width, and resolution of a network, resulting in increased performance with fewer parameters. This renders them particularly suitable in medical imaging processes that require precision and effectiveness. The results of the models depending on Efficient Net are briefly summarized in Table 3, and indicate the most efficient trade-off between the accuracy and the computational cost. The discussion points out that it is important to choose an optimal feature extraction backbone that can be used to obtain high prediction accuracy, and at the same time, be feasible to implement in clinical practice.

**Selection of EfficientNetB4 Backbone**

Following the comparative analysis, EfficientNetB4 was chosen as the backbone feature extractor because it has the best balance between model complexity and predictive performance(17). EfficientNetB4 is based on a compound scaling scheme; the depth, width, and input resolution of the network are scaled in the same direction; this guarantees that model capacity is utilised efficiently. It has some 19 million parameters and an input resolution of 300x300 pixels, making it sufficiently representative to be able to encode the detail of skeletal features and not being overly computationally intensive. Even compared with larger versions like EfficientNetB7, which demand much more computing power, EfficientNetB4 provides a feasible compromise between accuracy and efficiency(18). The model results in a 1792-dimensional feature vector that pools globally, which is basically radiographic characteristics, which are crucial in estimating the age of the bones. This feature representation forms the backbone of the multitask learning framework and allows classification and regression tasks. Furthermore, EfficientNetB4 can be generalized to a great number of datasets, implying that it can be used in clinical practice(19). It is also more acceptable in the deployment environments because of its support for moderate hardware needs. As such, the choice of EfficientNetB4 will guarantee that the model will perform well and be computationally efficient and scalable to real-world applications.

**Image Preprocessing Pipeline**

A clear preprocessing pipeline is employed to standardize input radiographic images and improve the performance of the models. It begins by decoding pictures in the general PNG and JPG formats and converting them to grayscale to emphasize the bone structures, but reducing the calculations. Images are then rescaled to a fixed 300 300 pixel bilinear interpolated, to ensure consistency in the data(20). The pixel intensity values are normalized to the range [0,1], which numerically makes them stable and their convergence to the training faster. To meet the EfficientNetB4 input specifications, the grayscale images are converted to three-channel representations. Also, channel-wise normalization of Efficient Net is obtained to align with the trained model parameters(21). The other preprocessing pipeline step is to include a batch dimension, which generates a model input-appropriate tensor. Notably, validation checks are also included so that only high-quality and relevant radiographic images are processed, decreasing the chances of making misleading predictions. The virtue of this preprocessing model can be linked to its systematic preprocessing approach that offers the model stability and predictability in feature extraction, which ultimately leads to superior predictive properties and reliability in clinical practices(22).

**Gender Encoding and Representation of inputs.**

The model is further modified with gender information as an additional input feature to increase the prediction and clinical relevance. The difference in the patterns of skeletal development of the male and female sexes is quite pronounced, and gender is thus a major determinant in establishing the age of the bones(23). Gender in this case is coded in binary form, with male taking the value of 1.0 and female of 0.0. Such a coded value is combined with the image-based feature vector, which is retrieved using the EfficientNetB4 backbone. In particular, using global average pooling, the model produces a 1792-dimensional feature that contains finer radiographic details. A one-dimensional gender input is then concatenated with this vector, and thus a combined feature representation of 1793 dimensions is obtained. The integrated feature vector is then fed to the multitask learning heads, which will carry out classification and regression tasks(24). The model will obtain a more detailed picture of skeletal maturity by incorporating both visual and demographic data. This methodology improves the precision of predictions and makes the system more similar to clinical assessment practice, when both the anatomy and the biological aspects are taken into account. Gender encoding inclusion is thus an important factor in enhancing the quality of models and their relevance in the real-life pediatric healthcare context.

| Method | Parameters | Input Size | Compound Scaling | Selected |
|---|---|---|---|---|
| VGG16 | ~138M | 224x224 | No | No |
| ResNet50 | ~25.6M | 224x224 | No | No |
| InceptionV3 | ~23.8M | 299x299 | No | No |
| MobileNetV2 | ~3.4M | 224x224 | No | No |
| EfficientNetB0 | ~5.3M | 224x224 | Yes | No |
| EfficientNetB4 | ~19M | 300x300 | Yes | YES |
| EfficientNetB7 | ~66M | 600x600 | Yes | No (overkill) |

**Table 3 Feature Extraction Method Comparison**

## EfficientNetB4-Based Bucket-Hybrid Architecture

The suggested methodology presents a new multitask deep learning network based on the EfficientNetB4 backbone, which will overcome the shortcomings of the traditional regression-based models. The architectural design is a bucket-hybrid approach, as the architecture is trained to learn both global skeletal and stage-specific developmental patterns. EfficientNetB4, which has the ability to scale in compounds, is used as the core feature extractor because it allows

learning radiographic features with high resolution efficiently. The input radiograph is fed into convolutional layers to produce a 1792-dimensional feature vector based on global average pooling. This representation of features provides high-resolution anatomical information to **Figure 3: Proposed EfficientNetB4 Bucket-Hybrid Architecture**
to make a precise age prediction of bones. The bucket-hybrid architecture further subdivides the learning process into predetermined developmental levels, enabling the model to make predictions into biologically significant age categories. It is a systematic method that makes the model more easily interpretable and compatible with clinical reasoning. The entire architecture combines common feature extraction and task-specific outputs, which allow effective multitask learning. Figure 3 demonstrates the detailed design of the proposed framework with the shared backbone and two-output architecture, with a focus on the interaction between pathways of feature extraction and regression.

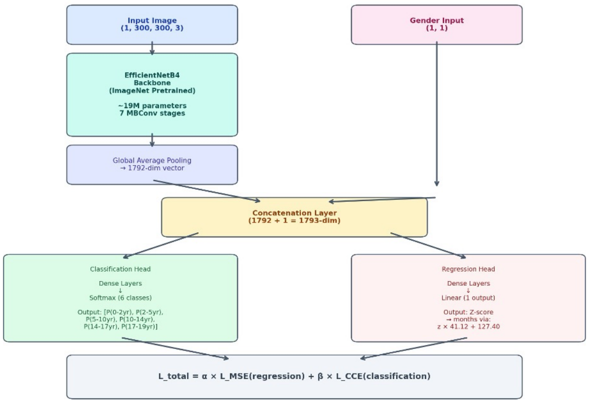

**Figure 3:** Proposed EfficientNetB4 Bucket-Hybrid Architecture

## Dual-Head: Learning Classification + Regression.

The model proposed uses a two-head learning strategy, which uses classification and regression tasks in a single framework. The classification head gives a guess of the bucket of the developmental stages of the input radiograph, and the regression head estimates the exact bone age value. The two-way (bilateral) task formulation helps the model utilize both the categorical and the continuous learning signals to enhance its accuracy and prediction strength(25). The classification branch, which consists of fully connected layers, takes the extracted features to a fixed set of developmental categories, and the regression branch will give a smooth bone age value, trained by representations. The model is an amalgamation of the most appropriate joint feature learning, which helps to reduce overfitting and enhance generalization. Further, the classification output offers contextual advice to the regression task, so that predictions do not go out of sync with biological stages of growth(26). This approach defeats the deficiency of single-task models, which tend to be non-interpretable and do not exhibit structured developmental patterns. The model is more believable and applicable to clinical use by classifying and regressing the skeletal maturity in a single architecture.

## Bucket Definition Strategy- Developmental.

In order to add biological context to the learning process, the proposed framework presents a definition strategy of a bucket of development of bone age that divides the bone age into stages with clinical significance. These buckets denote different stages of development of the skeleton, i.e., infancy, early childhood, mid childhood, adolescence, and near maturity. The definition of each bucket is given with the consideration of certain age ranges, and the model is able to get acquainted with the stage-wise peculiarities of bone development. This hierarchical classification assists the classification head in recognizing the stage of development of the input image, which then assists the regression head in making more precise predictions of age. The definitions of the buckets are well formulated to represent the clinical practice in the real world and to have a balanced representation of the various ages. The method decreases uncertainty in predictions and enhances the interpretability of models by making intermediate results that can be interpreted and verified by clinicians. Table 4 gives the detailed age ranges and respective bucket definitions used in the model. The developmental bucket strategy increases the accuracy and clinical applicability of the system by incorporating domain knowledge into the learning process.

| Bucket | Age Range | Label | Developmental Stage | Key Skeletal Features |
|---|---|---|---|---|
| 0 | 0-24 mo. | 0-2yr | Infancy | Carpal ossification begins |
| 1 | 24-60 mo. | 2-5yr | Early Childhood | Progressive carpal ossification |
| 2 | 60-120 mo. | 5-10yr | Mid Childhood | All carpals visible, sesamoid absent |
| 3 | 120-168 mo. | 10-14yr | Early Adolescence | Sesamoid appears, gender divergence |
| 4 | 168-204 mo. | 14-17yr | Late Adolescence | Progressive epiphyseal fusion |
| 5 | 204-228 mo. | 17-19yr | Near Maturity | Near-complete fusion |

**Table 4:** Developmental Bucket Definitions

### Multi-Task Loss Function Optimization

A joint loss function that minimizes classification and regression errors is used to optimize the multitask framework. Categorical cross-entropy loss is used to perform the classification task and is a loss that quantifies the difference between the predicted and actual buckets of development. Mean squared error (MSE) loss is used to optimize the regression task, so that the continuous bone age values are predicted accurately. The weighted summation of these two components of loss enables the model to weigh the significance of tasks in the training process. This collaborative approach to optimization allows this model to learn to complementary representations that enhance overall performance. The classification loss assists the model in learning the differences by stage, whereas the regression loss guarantees accurate age prediction. Through its combination of these two goals, the model generalizes better and minimizes the error in prediction in comparison to one-task methods. Also, the multitask loss functionality facilitates stable training and quicker convergence because it gives various learning indicators. Such an optimization method is essential to the achievement of high predictive accuracy and interpretability of the model; therefore, it can be used in clinical applications.

### Confidence Estimation Mechanism

To ensure better clinical reliability, the proposed system will include a mechanism of confidence estimation that determines the trustworthiness of model predictions. This is a mechanism that allocates confidence scores on the basis of the consistency of prediction and probability distributions of the classification head. Such scores are divided into various levels of confidence, including high, medium, and low, and they give more information to clinicians on the reliability of predictions. The confidence model aids in detecting uncertainty situations, which might need manual inspection, thus enhancing patient safety. The meaning of confidence levels is defined in a systematic manner to make it clear and practical in clinical practice. The close-up of the correlation between the score of confidence and the interpretation is shown in Table 5, which offers a systematic approach to the interpretation of the reliability of the prediction. The system goes beyond mere prediction and assists in making informed clinical decisions by incorporating confidence estimation in the model. This aspect makes a major contribution to the practical relevance of the model as it allows the deployment of the risk-sensitive system and increases the trust in automated systems to measure the age of the bones.

| Confidence Level | Bucket Probability | Clinical Interpretation | Action |
|---|---|---|---|
| High | >= 70% | Strong stage match | Routine screening |
| Medium | 40-69% | Transitional features | Standard review |
| Low | < 40% | Uncertain classification | Expert review required |

**Table 5:** Confidence Level Interpretation Framework.

### Framework and Safety Mechanisms of validation.

### Six-layer Radiographic Image Pipeline of validation.

The proposed system includes a six-layer radiographic image validation pipeline to make predictions reliable and safe. This pipeline aims to check the quality, authenticity, and appropriateness of the input images and then process them through the deep learning model. Every validation layer has a particular purpose, such as format checking, resolution checks, anatomy checks, and feature integrity checks. The pipeline automatically filters invalid or low-quality inputs, e.g., non-radiographic images or incorrectly captured scans, which would otherwise cause incorrect predictions. The system filters the images to the prediction model by applying strict validation criteria at various levels to pass clinically meaningful images. This systematic method makes it more robust and reduces the chances of giving inaccurate results. Detailed configuration of each validation layer along with the respective threshold is summarized in Table 6, and the overall workflow of the pipeline is as shown in Figure 4 which gives a clear view of the step-by-step process of validation.

| Layer | Check | Threshold | Failure Mode |
|---|---|---|---|
| 1 | Color Saturation (HSV) | Mean sat. > 50 | Photographs, color images |
| 2 | Pixel Intensity Variance | Std < 0.03 | Blank/corrupted files |
| 3 | Brightness Envelope | Mean > 0.95 | Overexposed images |
| 4 | Intensity Distribution | Peak conc. > 0.85 | Synthetic gradients |
| 5 | Dynamic Contrast Range | P95-P5 < 0.12 | Low-contrast copies |
| 6 | Edge Feature Density | Ratio not in [0.005,0.35] | Photos or featureless |

**Table 6:** Six-Layer Validation Pipeline with Thresholds

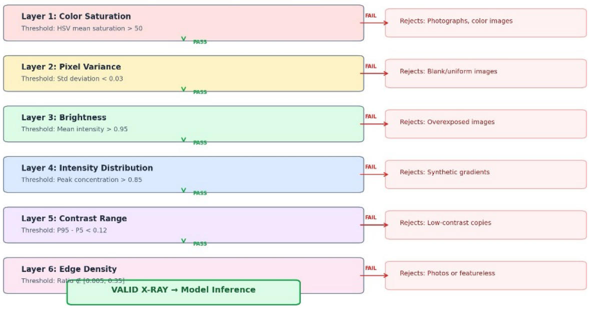

**Figure 4:** Six-Layer Radiographic Validation Pipeline

### Image Quality Assessment Using Thresholds.

Image quality assessment by a threshold is a mechanism that is essential to the validation mechanism to ascertain the suitability of radiographic inputs. The mechanism involves quantitative limits to measure key image attributes such as brightness, contrast, amount of noise, and sharpness of the structure. These thresholds will be well adjusted, so that only high-quality images that are of a clinical standard are carried on. Images that do not satisfy these requirements are automatically discarded or put on the manual review list to avoid inaccurate predictions. The threshold-based approach provides a standardized approach to managing quality, reducing variation, and ensuring consistency among different input samples. Moreover, the mechanism makes the model stronger by avoiding the impact of bad-quality data when making inferences. The system will ensure that the quality of images is highly controlled by integrating the use of quantitative assessment with the other validation pipeline. This approach is important in boosting accuracy and reliability in predictions, and the model can be more relevant in the actual clinical scenario, where variability in the input is a normal occurrence.

### Three-Tier Safety Layer Architecture

To further improve the reliability of the systems, an architecture of three levels of safety layers is proposed, which consists of another level of verification over the validation pipeline. Such is set to identify anomalies and make sure that predictions are only made of valid and relevant inputs. The initial level is concerned with the basic input verification, where the image should pass basic requirements like format or resolution. The second level carries out more sophisticated tests, such as anatomical checking and consistency checking of features to ensure that the input is a valid hand radiograph. The third level involves decision-level checking, where the predictive results are checked in terms of consistency and plausibility. Such a multi-layered safety system makes it harder to make erroneous predictions and can increase trust in systems. The hierarchical nature of the safety layers is arranged in such a way that errors that may occur are detected and countered at various levels of processing. This general structure of the safety framework is illustrated in Figure 5, which provides the hierarchical structure of the three levels of validation and their contribution to the strong and safe functioning of models.

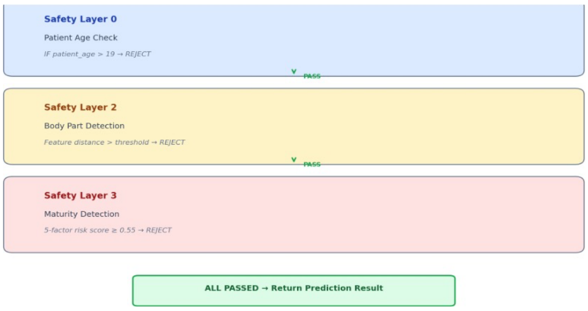

**Figure 5:** Three-Tier Safety Layer Architecture

**Figure 5:** Three-Tier Safety Layer Architecture

### Body Part Detection based on Feature Distance.

Another safety feature that will be adopted in the framework proposed is body part detection on the basis of feature distance analysis. This method makes sure that the input image is related to the region of the hand and wrist precisely, which is crucial when predicting the age of bones. The technique is to get feature representations of the input image and compare it against reference feature vectors obtained using valid hand radiographs. The system can ascertain whether the input image falls within the category of an expected anatomy by computing the difference between the feature representations of the same. In case the distance of the feature exceeds a predetermined threshold, the picture is invalidated and disqualified. This is a robust technique of identifying irrelevant or erroneous inputs, say images of other parts of the body or non-medical images. The system improves its capability of ensuring input integrity and avoiding faulty predictions by adding feature-based validation. This process is important in making sure that the model can work within the intended scope, thus enhancing accuracy and safety.

### Skeletal Maturity Detection Mechanism

A skeletal maturity detection mechanism is created to make sure that predictions are consistent with biologically plausible developmental stages. This component is a consistency measure, and it contrasts the predicted age of the bones with the structural features that are present in the radiograph. The system relies on the evidence of the reasonable range of the predicted age by assessing the patterns, such as the ossification centers and bone morphology, to ascertain the presence of the anticipated degree of skeletal maturity. The mechanism assists in identifying outliers or abnormal predictions that may take place due to unusual input conditions or inaccuracy in the models. It is also aiding in the estimation of the confidence framework, as it provides another test of reliability in prediction. The skeletal maturity detection renders the model easier to interpret since the model predictions can be supported by features that can be noticed in the clinical cases. This is to ensure that the system not only produces the right numerical results but also that it is consistent with the medical knowledge, thereby improving its applicability to clinical use and decision-making.

## System Deployment and Clinical Workflow Integration.

### End-to-End System Architecture

The suggested system is structured as a focus on a full-end architecture which includes the data-feed, preprocessing, model inference and visualization of the results in one system. The first step involves the uploading of pediatric hand radiographs through user interface, preprocessing and validation, which guarantees quality and accuracy of the image. The validated pictures are then sent to the EfficientNetB4-based multitask model, which does classification and regression to estimate bone age and stage of development.System architecture is designed in a way that allows flow of data between the components to be seamless in order to achieve efficient processing and minimum latency. All modules are designed in a modular way, which allows them to be flexible and scalable. The architecture also has validation and safety measures that would ensure that only relevant inputs are processed to enhance reliability. Frontend, backend, and model are smoothly integrated to provide a fluent exchange of information and real-time prediction opportunities. The entire system workflow, the data flow among modules and processing steps is shown in Figure 6 that gives a full picture of the architecture and makes it clear that it is quite suitable in clinical applications.

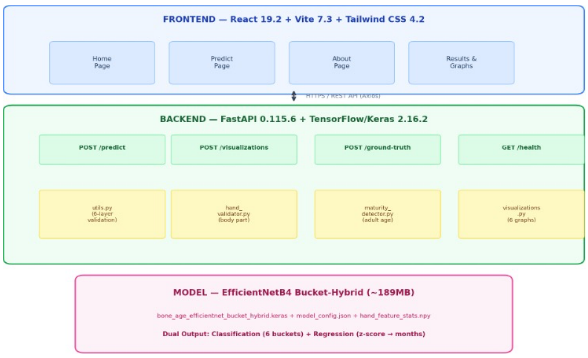

**Figure 6 Full System Architecture**

### Web-Based Clinical Deployment Framework.

The proposed system is also adopted as an online clinical application which can be easily accessed by the healthcare professionals to facilitate application in practice. The deployment model also allows radiographers to save the radiographic images on a web-based platform without necessarily installing special programs. The uploaded images are processed and provides bone age prediction and the levels of confidence and validation results in real time. This will enable it to be accessible and scalable, and that is why it will be easily compatible in hospital systems and diagnostic centers. The web-based design is also remotely usable with clinicians having the option of using the system elsewhere. Another goal of the framework is its capability to handle the large number of requests in an efficient way, and the stability of the performance in the various workloads. Security and data management are also taken into consideration to make sure that the data's confidentiality and compliance with clinical standards are maintained. The deployment structure helps fill the gap between the developed AI models and the real-world application since it provides a user-friendly and intuitive interface to ensure that the system could be easily integrated into the daily healthcare setup.

### Technology Stack and Implementation Details.

The suggested system implementation is based on scalable and effective technology stack which offers effective model deployment and system integration. The Python-based frameworks develop the backend that communicates with the model inference, data processing, and API communication. Current machine learning libraries assist in the deployment of deep learning model and achieve efficient training and inference. The frontend interface is developed with the help of the web technologies so as to make it an interactive and user-friendly interface. RESTful APIs are used to manage data flows between frontend and backend components to ensure a smooth flow of the data. Cloud-based or server-based deployment is also implemented in the system to increase the scalability and performance. The selected technology stack is compatible, flexible, and easy to maintain, which also makes it possible to adapt the system to various clinical settings. The exact components and tools to be adopted are as follows in the implementation, Table 7, which specifies the technologies applied in every level of the system. This systematic procedure will make the system technologically sound as well as practically implementable in real-life situations.

| Component | Technology | Version |
|---|---|---|
| Deep Learning | TensorFlow / Keras | 2.16.2 |
| Image Processing | OpenCV (headless) | 4.10.0 |
| Web API | FastAPI (async) | 0.115.6 |
| Numerical | NumPy | 1.26.4 |
| Visualization | Matplotlib + Seaborn | 3.9.2 / 0.13.2 |
| Frontend | React | 19.2.0 |
| Build Tool | Vite | 7.3.1 |
| CSS | Tailwind CSS | 4.2.0 |
| HTTP Client | Axios | 1.13.5 |
| Icons | Lucide React | 0.575.0 |

**Table 7 Technology Stack**

### Clinical Decision Support Visualization Engine.

The proposed system has one of the significant components in the visualization engine, which provides interpretable results to help in clinical decision-making. The engine delivers the findings of the prediction in a simple and well-structured form, including bone age (estimated), stage of development (classification), and confidence levels. Charts, indicators, and annotated outputs, as visual elements, are employed to improve readability and give clinicians valuable insights. This will enable health care professionals to easily determine the accuracy of predictions and make informed decisions. The backend model is connected with the visualization system, which displays the results in real time after their processing. Moreover, the engine outlines validation results, which enables users to identify issues with input pictures. It is more transparent and generates trust in automated predictions because it gives both numerical and visual data. Such a blend of visualization and prediction makes certain that the system not only give us valuable results, but it also allows effective interpretation of the clinical of bone age in the real-life healthcare setting and hence proves to be an effective tool insofar as assessment of bone age in pediatrics is concerned.

### Results and Performance Evaluation.

### Quantitative Performance Metrics.

The proposed multitask deep learning model was experimented on according to the traditional quantitative pointers to establish its accuracy and trustworthiness to anticipate bone age among children. The main measurement tool employed in this study is Mean Absolute Error (MAE), which is used to measure the mean difference between the predicted and actual bone age. The model has an MAE of approximately 9.10 months within the validation dataset which is competitive and is also easy to interpret and robust. In addition to MAE, the other performance measures, such as prediction consistency and validation accuracy, were also considered to ensure that they were all considered. The model can generalize across age groups and genders, which means that the model is effective as well. Validation and safety features can also assist in improving the reliability since only quality inputs will be put in. The particular quantitative results, including the data set size, the MAE values, and the circumstances of the assessment, are condensed in Table 8, and they provide a satisfactory view of the performance of the model on the validation data.

| Metric | Value | Unit |
|---|---|---|
| Mean Absolute Error (MAE) | 9.10 | months |
| Median Absolute Error | 7.07 | months |
| Model Inference Latency | 0.05-0.15 | seconds |
| End-to-End Processing | 0.15-0.45 | seconds |
| Input Resolution | 300 x 300 | pixels |
| Model Parameters | ~19M + heads | parameters |
| Model File Size | ~189 | MB |
| Validation Set Size | 1,262 | images |

**Table 8 Quantitative Results**

### Error Distribution and Model Accuracy.

The analysis of the error distributions gives more information on how the proposed model predicts in various samples. The findings reveal that most predictions are in a reasonable error margin, which proves that the model can be consistently accurate at different stages of development. Less pronounced errors are noted in intermediate age groups, where skeletal characteristics are more pronounced and minor deviations are possible in extreme age groups like infancy or late adolescence, because of minor anatomical differences. The error distributions of the predicted errors indicate that the model can give consistent results with a low variance, which is a necessary condition to be used in clinical practice. Moreover, multitask learning also enhances accuracy since it makes use of both classification and regression results. The finer resolution of errors in prediction, and the associated confidence levels with each prediction, is shown in Figure 7, which graphically illustrates the performance of the model over the validation dataset.

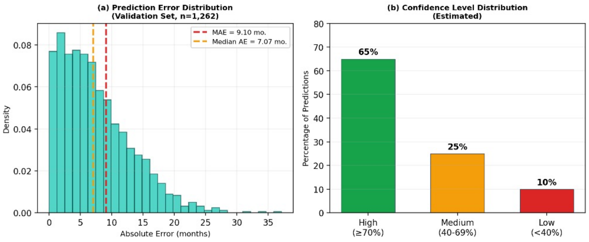

**Figure 7:** Prediction Error Distribution and Confidence Levels

### Confidence Level Distribution.

The proposed system will include a mechanism of confidence estimation that will give extra information about the trustworthiness of predictions. The levels of confidence are grouped into various ranges which enables clinicians to understand the degree of certainty with each prediction. The case of high-confidence predictions is associated with the situation when the model is characterized by high levels of agreement between classification and regression predictions whereas low levels of confidence suggest possible uncertainty or variability of predictions. The results of the distribution of the level of confidence on the validation dataset can be used to show that a large percentage of the predictions were in the high-confidence range, which is a good indication of the well-performing model. This characteristic improves the clinical usability of the features as it allows practitioners to recognize the cases that might need additional examination or confirmation. Confidence estimation integration is not only more transparent, but also helps to make decisions that are risk-aware. Using prediction results and confidence scores, the system would provide a more detailed analysis of results, thus making the automated bone age assessment systems more trustworthy.

### Comparison with Human Experts and Existing Models.

The proposed model was compared with human expert ratings and current deep learning models to determine its relative performance. The traditional methods such as the Greulich and Pyle method tend to be more varied with the reported values of MAE of about 7.32 months which corresponds to a variance in inter-observer. In contrast, more recent deep learning architectures have reported lower values of MAE, down to 5 months, using more advanced architectures and training strategies. The proposed model has an MAE of about 9.10 months, which is a bit higher than other state-of-the-art regression-based models. However, this difference can be attributed to the additional feature incorporated on the system, including multitask learning, validation pipelines, and confidence estimation systems. These properties enhance the interpretability, safety and clinical applicability, which is not typically talked about in conventional models. This analogy serves to highlight how necessary numerical accuracy is, but other qualities, such as reliability and usability, are also important in the real world.

### Clinical Utility vs Accuracy Trade-off.

The designed system is based on a trade-off that is balanced between the precision of predictions and clinical usefulness which is the primary drawback of the current models. Although it is possible to find deep learning methods with lower MAE, they are not always interpretable, do not have validation, or cannot be integrated into clinical practice. On the other hand, the proposed model is concerned with predictive performance in addition to reliability, safety and usability. **Figure 8:** Accuracy vs Clinical Utility Comparison The presence of validation pipelines means that only relevant inputs are operating, and the confidence estimation mechanism gives extra-context to deciphering predictions. Such holistic treatment makes the system more practical, although it might lead to slightly higher MAE than a system based purely on regression. The accuracy versus clinical utility is graphically represented in Figure 8, which indicates the trade-off that the proposed framework can provide. This discussion shows the system is more appropriate in a real clinical setting where reliability and interpretability are equally valuable as numerical precision.

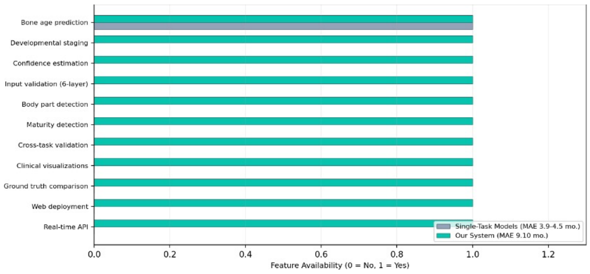

**Figure 8: Accuracy vs Clinical Utility Comparison**

## New contributions and innovations.

### Multi-Task Learning Framework

The proposed research paper introduces a new multitask learning model, which is the best to enhance bone age prediction and its comprehension among children. The model, unlike the traditional methods, integrates classification and regression tasks into a single architecture as opposed to regression only. Continuous bone age and categorical development stages can be learned at the same time using this multi-task approach The model can be used to forecast predictions which are consistent with biologically significant developmental stages, enhancing clinical applicability by classifying developmental buckets. The same feature extraction backbone allows efficient learning, with task-specific heads to optimize towards each objective. This method will not only enhance prediction consistency but will also minimize overfitting since shared representations are used. Also, the multitask model increases model interpretability, since clinicians can interpret predictions both in terms of precise age and development stage. This new technology assists in getting rid of one of the main shortcomings of the existing models, which can be described as the inability to structure the learning process and awareness of context. Overall, the multitask learning system is one of the significant innovations in the automated bone age assessment because of the combination of accuracy, interpretability, and clinical application.

### Integrated Confidence Estimation

The introduction of a credibility estimation system that examines the credibility of forecasts is one of the key innovations of the constructed system. This system is far superior to the traditional models because it only provides a numerical result; this one provides confidence levels to each prediction, as it allows clinicians to decide how reliable the results are. Confidence is estimated depending on the consistency of the prediction and probability distribution, such that the aspects of classification and regression are employed to arrive at the final score of confidence. The predictions can be coded into different levels of confidence, such as high, medium, and low, which provides a systematized explanation of the reliability of the model. This is particularly handy in clinical practice where there may be uncertain cases to be further analyzed or expertly analyzed. This system does not rely on external calibration methods, and the transparency it provides is simply to estimate the confidence of the model. This innovation increases the trust and helps to make informed decisions of the user, thereby contributing to the increased applicability of the system in the field. Confidence estimation is a key step towards the development of trustworthy and clinically acceptable AI systems.

### Robust Validation and Safety Pipeline

The proposed model will have an intensive validation and safety pipeline that will ensure integrity and reliability of input data and model predictions. There are several validation layers of this pipeline which determine quality of the image, anatomical accuracy and the appropriateness of the input before processing. The system removes the chances of inaccurate forecasts by selecting out invalid or low quality photos and enhances the overall resilience. The availability of threshold-based quality analysis and screening of body parts ensure that only the relevant radiographic images get processed. The architecture of three-layer safety layer, too, also, further supports the predictions in the evaluation of the consistency and plausibility of prediction. This is a broad approach to an important shortcoming of existing models, which in many cases do not have integral validation mechanisms. The safety checks at multiple points can be combined to be able to trust the system to perform reliably in the real-world clinical environments. This innovation has a great contribution to the durability and reliability of the model that can be deployed to healthcare facilities where precision and safety is of paramount importance.

### Clinical Visualization and Deployment.

The other valuable impact of the study is the creation of clinically oriented visualization and deployment framework which helps bridge the gap between AI models and the actual practices. The system offers an easy-to-use interface, enabling clinicians to submit radiographic images and get predictions in a format that is understandable. The visualization engine portrays some of its results as the approximate age of bone, stage of development and the level of confidence, which is measured by the well-defined graphical representations. This enhances the understanding and enables clinicians to make sound decisions based on model outcomes. The system is also implemented as a web-based application that guarantees accessibility and scalability to various clinical environments. The frontend, backend and model integration allow smooth functionality and real-time prediction. This end-to-end deployment model deals with one of the limitations of most of the research models that are not intended to be used in practice. The contributions of the proposed system in general, such as multitask learning, confidence estimation, validation mechanisms, and deployment, are summarized in Table 9, and they reflect how the research would fill the gaps in current research and further the research in the areas of automated bone age assessment.

| | Contribution | Gap Addressed | Prior Art Limitation |
|---|---|---|---|
| 1 | Multi-task learning (6-class + regression) | No dual outputs | Single-task only |
| 2 | Native softmax confidence (3-tier) | Requires MC dropout/ensembles | Added latency |
| 3 | Six-layer X-ray validation pipeline | No input quality assurance | Silent processing |
| 4 | Body part detection via own features | No body part check | Misleading results |
| 5 | Multi-factor maturity detection (5 factors) | Adult X-rays undetected | Out-of-range errors |
| 6 | Sex-aware dual-head modulation | Sex benefits a single output | Underexplored |
| 7 | Cross-task consistency validation | No self-checking | Errors undetected |
| 8 | Clinical visualization engine (6 graphs) | Bare number outputs | No visual context |
| 9 | End-to-end web deployment | Deployment neglected | Research-only proofs |
| 10 | Ground truth comparison (RSNA lookup) | No built-in validation | Cannot verify |

**Table 9 Novel Contributions**

### Discussion

The suggested multitask deep learning architecture proves to have a strong clinical value, proposing an automated and interpretable system to determine bone age in pediatrics. Through the combination of both classification and regression, the model matches the predictions to biological development phases, enhancing the understanding of diagnoses over traditional ones, like Greulich and Pyle and Tanner Whitehouse, which are not only subjective but also time-intensive. The proposed system has superior robustness to the current deep learning models, with in-built validation pipelines and confidence estimation, making predictions safer and more reliable. These characteristics render the system more applicable in the real-world clinical context, where the variability of input and reliability of decisions are essential issues. However, the research has certain limitations, including a small mean absolute error as compared to certain regression-only models and the use of a single data set that could affect the generalization to different populations. In addition, the computation requirements, though not intensive, could also be a bottleneck in the deployment to resource-constrained environments. Future studies are needed on how to enhance model accuracy with bigger and more diverse data, employing explainable AI methods to enhance interpretability, and real-time clinical validation. It can also be expanded to incorporate other diagnostic features to the system thereby making the system more applicable and aiding in the overall clinical decision-making.

## 12. Conclusion

The present paper presents an effective and clinically-oriented system of age prediction of children bones through multitask deep learning techniques. The proposed model is a powerful means of integrating the classification and regression in order to get the correct age estimation and the data regarding the stage of development to enhance the interpretability and clinical relevance. Extremely efficient feature extraction with EfficientNetB4, and potent preprocessing, validation piping, and confidence estimation engines, ensure strong and trustworthy performance. In contrast to traditional systems and most available deep learning models, the proposed system focuses on not just the predictive accuracy but also the safety, usability, and integration into clinical workflows. The model is marginally less accurate than some of the state-of-the-art methods, but it also makes a good trade-off between accuracy and the real-world considerations of deployment. Further facilitated by the development of a web-based system, accessibility and real-world application in healthcare settings can be achieved. Overall, the work contributes to reducing the gap between artificial intelligence research and its application to clinical practice by providing a scalable, interpretable and reliable tool to estimate the age of bones in children and be able to make more sound diagnostic judgments.

**References**

1. Yuan W, Fan P, Zhang L, Pan W, Zhang L. Bone age assessment using various medical imaging techniques enhanced by artificial intelligence. Diagnostics. 2025;15(3):257.
2. Marya A, Kapoor P, Kurniawan A. Methods of Skeletal Age Assessment. Dental Age Assessment: A Global Perspective: Springer; 2026. p. 293-321.
3. Kihara EN. Utilization and Accuracy of Dental Age Estimation Methods and Development of Tooth Maturity Reference Values in a Kenyan Population: COHES-JKUAT; 2025.
4. Longo UG, Villa Corta M, Valente F, Ruzzini L, D'hooghe P, Samuelsson K, et al. Skeletal Maturity Assessment in Pediatric ACL-Reconstruction. Children. 2025;12(9):1186.
5. Priyanka M, Manohar N, Suresha R, Dayananda R, Nair SSK, Sheela T. X-Ray Images for Automated Bone Age Estimation: Historical, Modern, and Prospective Patterns. IEEE Access. 2026;14:23367-94.
6. Oztel I. Ensemble Deep Learning Approach for Brain Tumor Classification Using Vision Transformer and Convolutional Neural Network. Advanced Intelligent Systems. 2025;7(10):2500393.
7. Ma C, Peng R, Li B, Zhang D, Zhang R, Zhang Z, et al. Application of multi-scale feature extraction and explainable machine learning in chest x-ray position evaluation within an integrated learning framework. European Radiology. 2025:1-15.
8. Hsu W-S, Lin G-T, Wang W-H. Real-Time Ultrasound Diagnosis of Developmental Dysplasia of the Hip Using an Attention-Enhanced YOLOv11 Model. International Journal of Medical Sciences. 2025;22(16):4236.
9. Rodriguez-Marin M, Orozco-Alatorre LG. Advancing pediatric growth assessment with machine learning: overcoming challenges in early diagnosis and monitoring. Children. 2025;12(3):317.
10. Alekseevich KL. Adaptive learning methods based on crowdsourced data labelling for medical images data mining.
11. Gulhane A, Kadam P, Warjurkar N. Automated Detection Of Rotten Fruits And Vegetables Using Deep Learning. 2025.
12. Chandrashekar G. Collaborative Deep Reinforcement Framework for Multimodality Integration and Learning: University of Missouri-Kansas City; 2025.
13. Shah P, Kumar Y, Bhowmik T. Deep Learning Techniques for the Diagnosis and Detection of Orthopedic Conditions: A Systematic Review of Recent Advances and Challenges. Archives of Computational Methods in Engineering. 2025:1-27.
14. Andleeb I, Hussain BZ, Joncas J, Barchi S, Roy-Beaudry M, Parent S, et al. Automatic evaluation of bone age using hand radiographs and pancorporal radiographs in adolescent idiopathic scoliosis. Diagnostics. 2025;15(4):452.
15. Çatak C, Ocak M, Çolak D. A two-stage deep learning approach for forensic age and sex determination from hand-wrist radiographs: an anatomical pose estimation and multi-task model. International Journal of Legal Medicine. 2026:1-18.
16. Kashif A, Shafique H, Khursheed A, Latif A, Ali S, Batool SN, et al. FORENSIC RADIOLOGY: AN INTELLIGENT METHOD OF AGE AND GENDER ESTIMATION FROM X-RAY SCANNED BONES IMAGES USING CONVOLUTIONAL NEURAL NETWORKS. Spectrum of Engineering Sciences. 2025:440-87.
17. Ghosh M, Basu B, Nandi A. Diagnosis of breast malignancy with microwave breast imaging method and utilizing EfficientNet-B4 deep learning model. IEEE Transactions on Instrumentation and Measurement. 2025.

18. Inbalatha M, Raghu N. Optimized Deep Learning Architecture for Pediatric Pneumonia Diagnosis in Chest Radiographs with Integration of EfficientNetB4, Topological Convolutional Layers, and Advanced Augmentation Strategies. Journal of Robotics and Control (JRC). 2025;6(4):1568-80.

19. Batool SN, Yang J, Gilanie G, Latif A, Yasin S, Ikram A, et al. Forensic Radiology: A robust approach to biological profile estimation from bone image analysis using deep learning. Biomedical Signal Processing and Control. 2025;105:107661.

20. Thelkar AR. Fundamentals of Digital Image Processing in Medical Applications: Academic Guru Publishing House; 2025.

21. Mishra VP, Roy A. Image processing strategy and target device. Advances in Image Processing, Reliability, and Artificial Intelligence: Elsevier; 2026. p. 209-23.

22. Paraskevas K, Christos T. Data preprocessing and feature engineering for data mining: Techniques, tools, and best practices. Ai. 2025;6(10):257.

23. Martín Pérez SE, Martín Pérez IM, Molina Suárez R, Vega González JM, García Hernández AM. The validation of the Tanner–Whitehouse 3 method for radiological bone assessments in a pediatric population from the Canary Islands. Osteology. 2025;5(1):6.

24. Riem L, Pinette M, DuCharme O, Pabon V, Morris J, Coggins A, et al. Big bones mean big muscles: an MRI-based dataset of muscle-bone-body size relationships across 70 human muscles of the upper limb, trunk, and lower limb. Journal of Applied Physiology. 2026;140(1):1-19.

25. Terblanche J. Linear and non-linear methods for the prediction of mandible morphology: Stelenbosch: Stellenbosch University; 2025.

26. Abdullah, Fatima Z, Sánchez Mejorada CG, Ather MA, Oropeza Rodríguez JL, Sidorov G. Fair and Explainable Multitask Deep Learning on Synthetic Endocrine Trajectories for Real-Time Prediction of Stress, Performance, and Neuroendocrine States. Computers. 2025;14(12):515.