International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Deep Learning Framework for Skeletal Age Classification from Pelvic Radiographs using K-fold Cross Validation and Stacking of CNN Models

S. Jenifer[1*], Dr. M. Arun[2]

Submitted: 05/01/2025 **Revised:** 25/02/2025 **Accepted:** 08/03/2025

Abstract: Recent technological developments in deep learning environments have improved bone age evaluation, making it easier and more exact than classic methods in forensic radiology. Deep convolutional neural networks are highly effective at detecting bone age; however, their complexity arises from the number of parameters they need, making them resource-intensive to run on CPUs. To address this, the proposed work utilizes the transfer learning approach to build a two-stage deep learning model based on pelvic radiographs, comprising a vital bone extraction model and an age assessment model. Initially, UNet model combined with Attention Gate extracted the pelvic girdle bones by filtering insignificant regions from pelvis X-rays. For age assessment, a smaller classifier network was first developed and evaluated using K-fold cross-validation. Subsequently, the two deep networks were built by layering the new ones to the existing framework. To enhance performance further, the outputs of both the classifiers were stacked using a dense layer called an aggregator. This meta-learner combined the strength of each model to make decisions on final prediction. The whole framework was validated to analyse its ability to categorize human age in the range of 0–19 years using the collected pelvic radiographs and achieved an average classification accuracy of 97.50%, precision of 98.25%, recall of 96.65%, and F1-score of 97.20%. Thus, the proposed framework can increase the accuracy of multi-classification tasks while leveraging the limited computational resources.

Keywords: *Bone age, Deep learning, Ensemble, Fine-tuning, K-fold cross-validation, Pelvic X-rays*

1. Introduction

Bone age is a key indicator to assess the growth and maturity of an individual's skeletal structure. It plays a vital role in legal proceedings involving minors and diagnosing growth anomalies in medical practices. In digital forensics, the radiography-based method is the simplest and most cost-effective way to ascertain age. While hand bones are extensively employed to measure bone ageing, pelvic bones can help better characterize bone maturity during adolescence [1]. The research conducted by [2] examined the pelvic computed tomography (CT) images in persons aged 8 to 16. Their research showed that assessing human age in forensic radiology can be done by looking at the ischiopubic and iliopubic synchondrosis of the pelvis.

In the context of bone age evaluation, plenty of studies were conducted to assess the different pelvic regions from the pelvic radiographs. The evaluation model [3] intended to analyse the Hounsfield Unit (HU) value from the os coxae areas of pelvic CT images for age determination, whereas method [4] used a grading system to analyse changes in the pubic symphyseal surface based on age, utilizing pelvic CT scans. Meanwhile, age detection approach [5] first classified the CT scans according to the Risser stages of iliac crest ossification, which ranged from 0 to 7. Subsequently, ten different regression models were established to estimate age using Risser phases. Likewise, the studies [6,7] explored the conventional regression approaches to predict an individual age based on statistical metrics of various pelvic bone regions.

In our study, we opted for pelvic X-rays over CT scans because they are more affordable and provide a better visual representation of bone structures. Also, age calculation using traditional machine learning methods required extensive experience to derive the statistical features of pelvic regions from radiographs. This problem was eventually fixed by deep learning networks, which can automate the feature extraction process to distinguish across classes.

In this research, we presented a deep learning-based structure for skeletal age determination from pelvic radiographs featuring segmentation and classification. We built the classification network architecture with the pre-trained auxiliary classifier to facilitate better feature acquisition. The following sums up the primary facets of this work:

- We demonstrated the bone extraction network for segregating the hand portion from the radiographs with the fewest labelled X-ray images.
- We exploited the potential of the K-fold cross-validation mechanism to assess the base model's reliability with a

1Research Scholar, School of Computing, Department of Computer Science and Engineering, Vel Tech Rangarajan DR. Sagunthala R and D Institute of Science and Technology, Chennai-600062, Tamil Nadu, India.
Email: jeniss37@gmail.com (Corresponding Author)*
2Professor, School of Computing, Department of Computer Science and Engineering, Vel Tech Rangarajan DR. Sagunthala R and D Institute of Science and Technology, Chennai-600062, Tamil Nadu, India.
Email: saiarun2006@gmail.com

minimal amount of data.

- We devised the two distinct classification network architectures utilizing pre-trained weights from the base model for calculating the bone age.
- We ensembled the outcomes from both the classifiers for accurate age categorization.
- We carried out experiments to establish the competitiveness of the proposed strategy against various baseline techniques.

The format of this paper is as follows: The subsequent section examines the existing studies pertaining to the research topic. Section 3 presents the deep learning-based structure for skeletal age determination featuring segmentation and classification. Section 4 depicts the research findings to establish the competitiveness of the proposed strategy against various baseline techniques. The final section highlights the research findings and suggests areas for further improvement.

2. Related Works

Numerous deep-learning approaches have been stated to find the human age by evaluating skeletal maturity from pelvic imaging. The study by [8] experimented with the AlexNet to find out bone age from pelvic radiographs. It was claimed that the improved AlexNet outperformed the current cubic regression model. Similarly, deep learning framework [9] demonstrated the pre-trained Xception model for age detection from pelvis images, whereas the deep CNN network [10] trained three pre-trained frameworks, VGG19, Inception-V3, and Inception-ResNet-V2 for skeletal age prediction using pelvic X-rays.

Furthermore, cross-validation (CV) is a useful technique for avoiding overoptimism in deep learning models. In research [11], the authors provided guidelines for choosing a suitable cross-validation strategy to process medical images. Different cross-validation strategies are employed to identify the top-performing machine learning algorithms for cervical cancer detection [12], breast cancer classification [13], early intrauterine foetal death detection [14], and identification of maternity risk level [15]. Besides, the automated framework [16] adopted the 5-fold CV technique to evaluate the CNN model for classifying brain tumours using magnetic resonance imaging. Therefore, it was clear from recent studies that using the CV approach for evaluating deep learning models enhances the system's resilience.

The deep learning framework [17] employed pre-trained CNN networks, which were fine-tuned by including dense layers for age prediction from hand radiographs. They subsequently experimented with different optimizers to assess each model for performance analysis. Besides, the research work [18] carried out an extensive review of various fine-tuning strategies for three well-known pre-trained networks: ResNet-50, DenseNet-121, and VGG-19, across different medical imaging fields. Their findings emphasized that the distinct features of medical images require careful consideration in architectural design and fine-tuning methods. Therefore, we decided to deepen the model's comprehension of the features in the training dataset by fine-tuning the model on the same data.

In conclusion, the existing methodologies demonstrated that CNN-based approaches were widely applied to estimate skeletal age from pelvic radiographs. This study employed pelvic X-rays to estimate ages, focusing on the pelvic girdle region. The proposed automated skeletal age estimation framework was devised using Attention UNet to isolate the pelvic girdle bones and fed the extracted bone regions into the deep CNN model for bone age evaluation. The experiments were carried out to emphasize the importance of incorporating a small, well-trained classifier into the CNN structure to enhance its performance.

3. Methodology

The proposed work endeavours to optimize the deep learning model for human age categorization while leveraging the limited computational resources. The purpose of this paper is threefold: (1) to calculate skeletal age with optimal accuracy; (2) to identify the best CNN model that validates across various subsets of the training samples; and (3) to enhance the predictive ability of the CNN model through weight transfer for unknown samples. The structure of the proposed age classification framework comprised three phases, as shown in Fig.1: Data preparation came first, followed by area of interest segmentation, and age evaluation network.

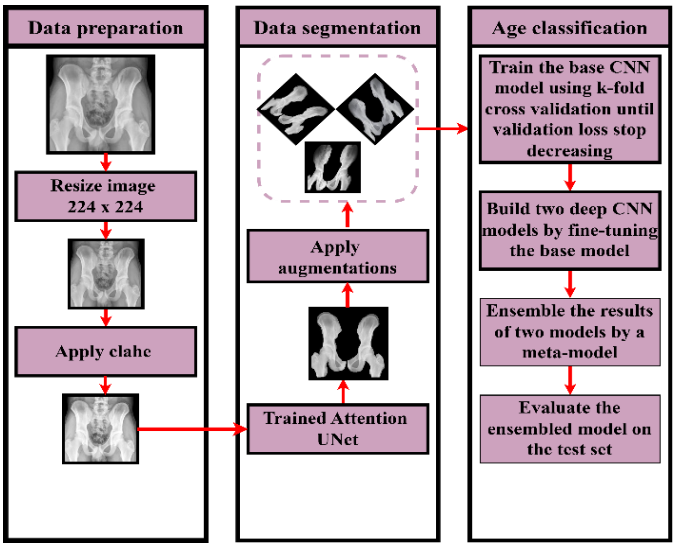

Fig. 1. The proposed framework for skeletal age evaluation

3.1. Data Collection

The dataset included 2520 conventional pelvic X-rays aged 0 to 19 years. A pelvic X-ray revealed the following structures: iliac crests, pelvic girdle, pubic symphysis, and proximal femora. Merely normal and healthy radiographs with no pelvic trauma or injury were considered for further evaluation. Fig.2 depicts some pelvic radiographs from the collection.

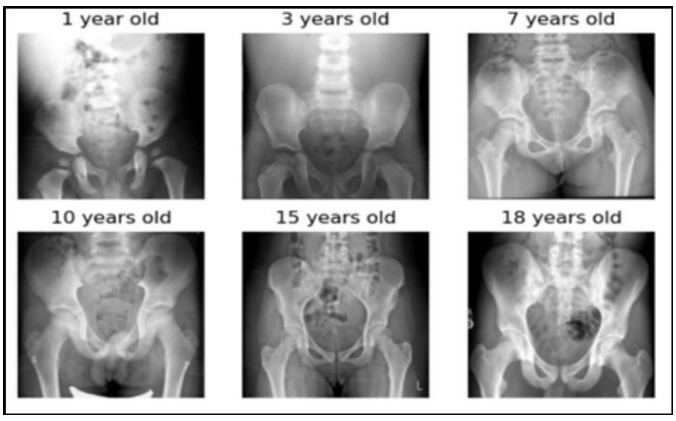

Fig. 2. Sample pelvic X-rays from the collected data

3.2. Image Preprocessing

Image pretreatment was done as follows to standardize images across the data collection: Resize the x-rays with a size of 256×256 pixels; remove noise from the images; enhance image contrast with the Contrast Limited Adaptive Histogram Equalization (CLAHE) technique.

3.3. Pelvic Bone Extraction Network

The proposed bone age evaluation framework estimated skeletal age by focusing on the ossification and union of the three major pelvic bones, namely the ilium, ischium, and pubis. Segmentation network is to locate the pelvic bones of interest for further evaluation. During the training process, binary images known as masks that specify the area of interest in an image, along with actual images, were fed into the extraction network as input. Eventually, the segmented dataset, which represents the masks, were generated.

The extraction network adopted in this work adopted a framework akin to UNet [19]. The contracting path of the network was made up of four convolutional modules, each with two convolutional layers. The sequence of convolutional blocks was as follows: two 3×3 convolutions with same padding, batch normalization, and Rectified Linear Unit (ReLU). Following each block, a max-pooling layer was placed to reduce the input size. Similarly, the bottleneck layer linking the encoding and decoding paths had two 3×3 convolutions.

In the decoding phase, the upsampling operation was carried out to double the image size. The attention module [20] was placed between the encoding module and the upsampling block to emphasis on the salient features at each scale. Next, the attention gate output was fused with the corresponding upsampled feature maps from the decoder block. Subsequently, two 3x3 convolutions were applied, accompanied by a ReLU function. Following a 1x1 convolution, the final decoder produced the segmentation mask with the dimensions of 256x256x1.

Table 1 provides information on the extraction model. During training, model's parameters were updated using Adam optimizer and learning rate was chosen as 0.01. The extraction model loss was estimated using the Dice loss function. Equation (1) represents the Dice loss calculation, where $mask_{pred}$ represents the output of the extraction network, and $mask_{true}$ is the labelled image.

$$\text{Diceloss} = 1 - 2 \times \frac{|\text{mask}_{\text{pred}} \cap \text{mask}_{\text{true}}|}{|\text{mask}_{\text{pred}}| + |\text{mask}_{\text{true}}|} \quad (1)$$

Table 1. Structural elements of the extraction model

	No of kernels	*Kernel size*	*Strides*	*Output*
Encoder_block-1	64	3 x 3	1	128x128 x64
Encoder_block-2	128	3 x 3	1	64x64x128
Encoder_block-3	256	3 x 3	1	32x32x256
Encoder_block-4	512	3 x 3	1	16 x16x512
Bottleneck	1024	3 x 3	1	16x16x1024
Decoder_block-1	512	3 x 3	1	32x32x512
Decoder_block-2	256	3 x 3	1	64x64x256
Decoder_block-3	128	3 x 3	1	128x128x128
Decoder_block-4	64	3 x 3	1	256x256x64
Final_conv	64	1 x 1	1	256x256x1

For training purposes, 1000 images were chosen for manual mask annotation. The rest of the images were segmented using the trained model. Next, we obtained the pelvic bone area by overlaying the extraction mask on the real image.

3.4. Age Evaluation Network

The segmented dataset was augmented by performing various transformations to enhance its diversity. Eventually, the transformed images were fed into the CNN-based model for skeletal age assessment. The supervised age classification network extracted the informative characteristics from the pelvic images and categorized them into 20 classes of age ranging from 0 to 19. The framework for calculating bone age was constituted by the following procedures:

1. Performing iterative k-fold cross-validation to determine the optimal classifier network.
2. Developing two deep CNN models and feed them the weights of the ideal base classifier.
3. Stacking the output of the both the models to ensure robust performance.

3.4.1. Structure of the Base CNN Model

In this part, we strived to transfer knowledge from a small CNN to a larger one through weights. Typically, weight distillation entails gaining knowledge from the weights of a deeper network. However, as previously indicated, we sought to achieve efficiency while implementing the deep learning model on low-cost hardware. Besides, the k-fold cross-validation approach enabled us to generalize the smaller network for the validation data.

The developed compact CNN classifier model had four convolutional layers, each succeeded by a max-pooling layer. In the initial layer, there were 16 filters, and this number doubled in the subsequent layers. Then came three dense layers, the last of which was the SoftMax layer, which divided the data into twenty age groups. The framework of constructed model is illustrated in Fig.3. To train the model, we relied on the k-fold cross-validation method, considering the model's size to prevent overfitting.

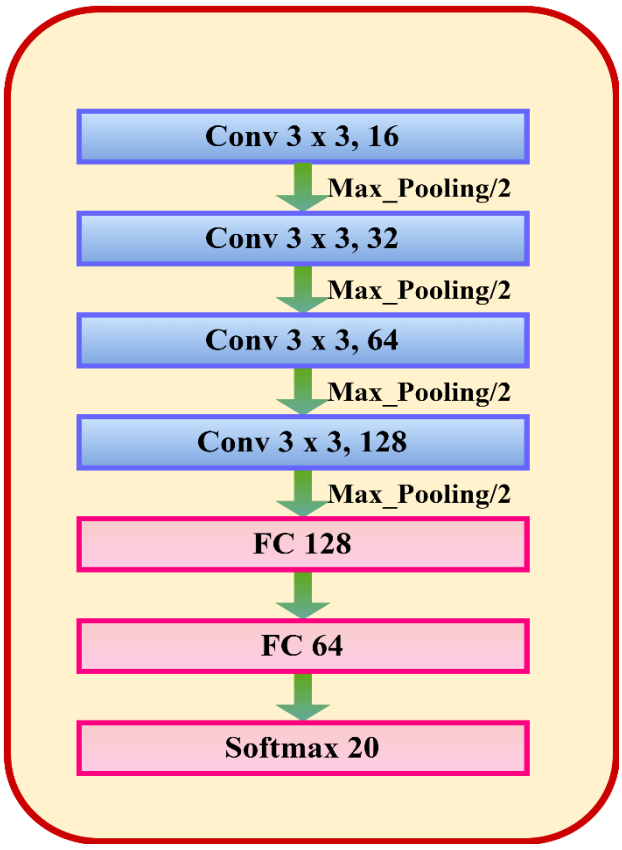

Fig. 3. The structural diagram of proposed CNN

Remember that the purpose of this method is to assess the model architecture rather than specific training since the same model was re-trained using various training sets. During model evaluation, the given data were split into k subsets, with one of these subsets as a validation set and training the model on the remaining subsets. In each of the k iterations, a different fold served as the validation set. This evaluation process was iterated n times, with the training samples being scrambled before each repeat, resulting in a new split of the samples.

3.4.2. Iterative K-fold Cross Validation

This section described the CNN model's training procedure as algorithm 1. The pelvic region extracted from the pelvis X-rays, measuring 224x224x3, was input into this algorithm. Subsequently, the training samples were segmented into k subsets. The model was then evaluated k times, using each fold as the test set once. During the iterative process, the model with the lowest validation loss was chosen at each iteration. The knowledge of this model was then passed to the next iteration through its weights. This procedure was continued until the validation loss ceased to decrease. At that point, the weights of the best-performing model were saved.

Algorithm 1: Model training procedure

1: **Input:** Extracted pelvic bone images X ϵ $\mathbb{R}^{L\times L\times C}$
2: **Initialization:** Define k to divide the training data into number of subsets
3: **do:**
4: **for** i= 1 to k:
5: Create model and load the pre_trained weights into it from the model[iteration-1]
6: Use (k-1) folds for training
7: Evaluate the model on validation fold
8: **if** validation_loss < best_loss
9: best_loss = validation_loss
10: **end if**
11: **end for**
12: Get model with the best_loss and restore the model weights
13: **until** best_loss [iteration] < best_loss [iteration-1]
14: **Output:** Moel weights from the final iteration

In the aforementioned cross-validation procedure, the model was initialized using weights from the previous iteration. The learning rate for the initial iteration was chosen as 0.01. Afterward, a meagre learning rate was applied, allowing subtle adjustments to the previously learned features while preventing the model from overfitting.

3.4.3. knowledge transfer

In this phase, we aimed to strengthen the age determination model's capacity to adapt effectively to unseen pelvis X-rays by refining the pre-trained model. To achieve this, we tried to develop two refined models and coupled them to capture various facets of the data, producing more accurate outcomes. In model 1, two more blocks were incorporated into the underlying model, each including two convolutional layers with a depth of 256,512, respectively, to refine the higher-level representations of the model. By allowing these new layers to be trainable, the model excelled on new samples while leveraging the information gleaned from the low-level layers within a related domain. Model 2 had an identical structure to model 1, except for replacing the base model's first two earlier layers with a single convolutional block that had two convolutional layers, each with 32 filters. Only the higher-level weights from the base model were carried over to the new model.Fig.4 describes the steps involved in the fine-tuning of the age determination model.

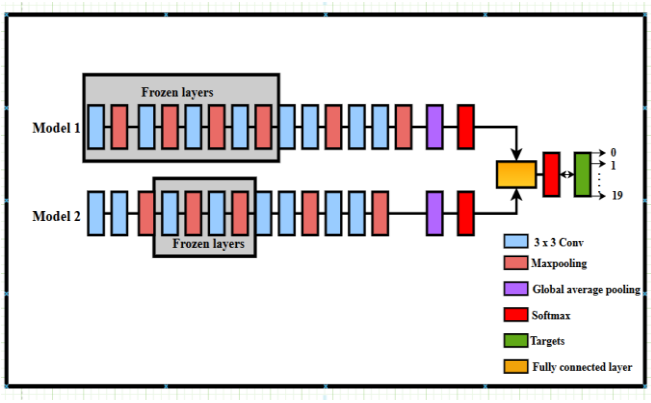

Fig. 4. Fine-tuning and training of age evaluation framework

The final step involved feeding the feature maps from the last module into a global average pooling layer to avoid overfitting, then into a SoftMax layer to categorize the data into twenty classes of age for both the models. Ultimately, a meta-model was established, comprising a fully connected layer of 32 neurons that consolidated the outcomes from the two base models.

3.4.4. Training of the ensembled model

During training, we considered the base classifier's knowledge because k-fold cross-validation provided reliable assessments of model performance. Thus, the age evaluation models utilized the weights from the best model of the final iteration of cross-validation. In the deeper models, the weights for the earlier layers from the base classifier were set to non-trainable to capitalize on the features taught from the pre-trained model. In the end, the predictions from both models were input to the meta-model which was then trained using original target values as true labels. Simultaneously, the performance of the meta-learner was assessed with a distinct test set.

The proposed approach relied on iterative learning to optimize the model parameters. The dataset was split at random into two sections: 80% of the radiographs were considered for training and 20% for validation. For parameter optimization, Adam optimizer was employed, and the batch size was fixed at 32. The Sparse categorical cross-entropy was the loss function that needs to be minimized. Equation (2) defines the loss function, where m is the number of classes, a_n is the bone age, and p_n is the probability for n^{th} class.

$$\text{Loss} = -\Sigma_{n=0}^{m} a_n \log(p_n) \qquad (2)$$

If there was no reduction in valid loss after five consecutive epochs, the training process was halted. The best-performing model and its parameters were retained.

4. Results

This section demonstrates the performance of a small classifier network using k-fold cross-validation on the pelvis X-ray dataset and the effectiveness of the recommended ensembled approach on test pelvic images while comparing it to other models.

4.1. Performance of the K-fold Cross-Validation

In this subsection, we assessed the robustness of the simple CNN classifier by training it on different subsets of the dataset. This process was iterated until the validation loss showed no significant decline. Table 2 summarises the training configurations for the k-fold approach.

Table 2. The training configurations for iterative k-fold approach

Parameters	*Value*
K	4
Batch size	32
Epochs	150
Optimizer	Adam
Learning rate for the first iteration	0.001
Learning rate for the remaining iteration	0.0001

After the fourth iteration, our model's validation loss failed to improve; therefore, we selected the model with the lowest loss from that iteration as the final model. Table 3 compares the validation loss from each iteration. Every iteration witnessed a decrease in validation loss, which allowed the model to converge more quickly as the best weights were retained each time.

Table 3. Validation loss from all the four iterations

Iteration	*Validation loss*
1	1.17
2	0.40
3	0.09
4	0.005

4.2. Evaluation of the Ensembled Model

The two stated deeper models, designated as model 1 and model 2, were trained using the same pelvis dataset. Figures 5, and 6 demonstrate the accuracy and loss over training epochs for model 1, and model 2, respectively. While model 2 outperformed model 1, it took 70 epochs to converge, whereas model 1 required only 28 epochs. Table 4 summarizes the accuracy of the proposed models.

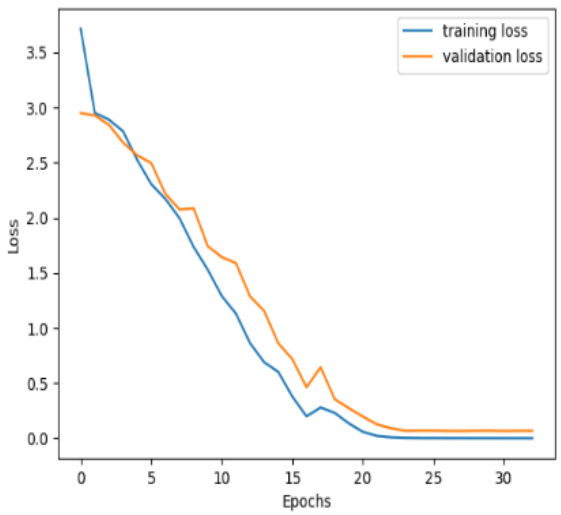

Fig. 5. Plot on loss over training epochs for model 1

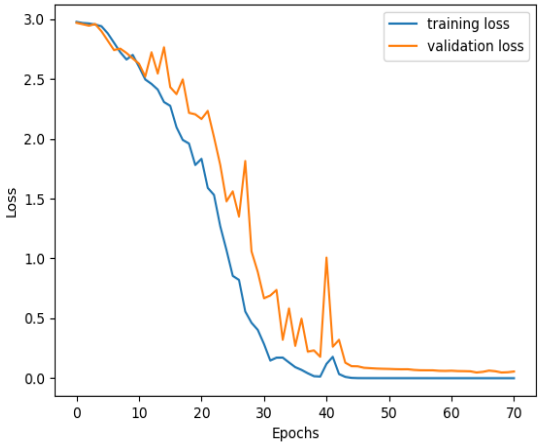

Fig. 6. Plot on loss over training epochs for model 2

Table 4. Accuracy comparison between the three models.

	Training accuracy	*Epochs taken*
Model 1	97.88	28
Model 2	98.12	70
Ensembled model	98.75	65

Evaluation metrics, namely precision, recall, and F1-score, were employed to review the performance of the suggested models. Precision is the model's capacity to not misclassify negative samples as positive for the target class. The recall (sensitivity) assesses the model's ability to identify relevant samples for the target class. The F1-score is the harmonic mean of precision and sensitivity. Tables 5 and 6 display the classification metrics for model 1 and model 2, highlighting the classes where each model performed well based on the validation data.

Table 5. Performance measures of model 1 for all 20 classes

Class (age in years)	*Accuracy*	*Precision*	*Recall*	*F1-score*
0	1.00	1.00	1.00	1.00
1	1.00	1.00	1.00	1.00
2	0.93	1.00	0.93	0.96
3	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00
5	1.00	0.86	1.00	0.92
6	1.00	1.00	1.00	1.00
7	1.00	0.91	1.00	0.95
8	1.00	1.00	1.00	1.00
9	1.00	0.94	1.00	0.97
10	0.82	1.00	0.82	0.90

Table 5. (cont.)

Class (age in years)	*Accuracy*	*Precision*	*Recall*	*F1-score*
11	1.00	1.00	1.00	1.00
12	0.60	0.75	0.60	0.67
13	1.00	0.67	1.00	0.80
14	0.70	1.00	0.70	0.82
15	1.00	0.57	1.00	0.73
16	1.00	1.00	1.00	1.00
17	1.00	1.00	1.00	1.00
18	1.00	1.00	1.00	1.00
19	1.00	1.00	1.00	1.00

Table 6. Performance measures of model 2 for all 20 classes

Class (age in years)	*Accuracy*	*Precision*	*Recall*	*F1-score*
0	1.00	1.00	1.00	1.00
1	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	0.96
3	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00
6	0.91	0.91	0.91	0.91
7	1.00	0.91	1.00	0.95
8	1.00	1.00	1.00	1.00
9	1.00	1.00	1.00	1.00
10	0.91	1.00	0.91	0.95
11	1.00	1.00	1.00	1.00
12	1.00	1.00	1.00	1.00
13	1.00	1.00	1.00	1.00
14	0.70	1.00	0.70	0.82
15	1.00	0.57	1.00	0.73
16	1.00	1.00	1.00	1.00
17	1.00	1.00	1.00	1.00
18	1.00	1.00	1.00	1.00
19	1.00	1.00	1.00	1.00

Table 6 indicates that model 2 performed slightly better than model 1, achieving an accuracy of 96.12%, precision of 97%, recall of 98%, and F1-score of 97%. Despite the performance disparity, model 1 predicted class 6 with 100% accuracy against 91% for model 2. Thus, it can be concluded that the ensemble technique utilized the strengths of various models for improved generalization.

Ultimately, the predictions from both models were fed into a neural network, which generated the final outcome. The confusion matrix is a table to illustrate the classification results by comparing the actual label with those predicted by the model. The diagonal elements in confusion matrix denote the number of valid predictions for each class.Fig.7 depicts the confusion matrix of the ensembled model following the evaluation of 160 test pelvic images.

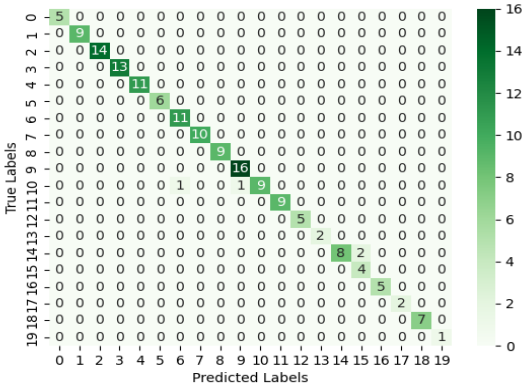

Fig. 7. Confusion matrix of the ensembled model

4.3. Discussion

This study relied on the following anatomical structures in pelvic radiographs to determine age: iliac crests, ilium, ischium, pubis, and proximal femora. The human pelvis may continue to grow until the age of 21. Nevertheless, the pelvic growth pattern differs with age during adolescence. At first, ischium and pubis begin to fuse to develop the ischiopubic region between the ages of 4 and 8. Subsequently, between the ages of 11 and 14, the ilium unites with the ischiopubic region. The iliac crest is a convex curve that defines the upper border of the ilium. As the person reaches adulthood, it starts to ossify in the centre of the crest and extends to the spine. In females, the iliac crest merger with the ilium begins at age 15, while in males, it commences at age 17.

We built a framework for assessing skeletal age by considering the growth pattern of the pelvic girdle and iliac crest epiphysis from pelvis radiographs. It had been found that the suggested ensembled model performed well even with a smaller dataset, with a mean accuracy of 97.50%.

Furthermore, the suggested framework outperformed the current CNN models while predicting the bone age on the same dataset. Table 7 compares the resultant performance of various CNN algorithms for age estimation. As stated in table 7, the average accuracy, precision, recall, and F1-score of the given framework were 97.50%, 98.25%, 96.65%, and 97.2%, respectively.

Table 7. Performance of various CNN algorithms for age evaluation

Model name	*Accuracy*	*Precision*	*Recall*	*F1-score*
VGG16	92.19%	90.65%	89%	89.81%
VGG19	93.75%	93%	92%	92.50%
DenseNet121	94.38%	94.25%	93%	93.6%
ResNet50	96.88%	95.25%	93.5%	93.6%
Proposed ensembled model	**97.50%**	**98.25%**	**96.65%**	**97.2%**

The experimental outcomes demonstrated that combining the two models initialized with well-trained weights can considerably enhance the categorization ability of the framework. It's necessary to know that the global age limits for criminal culpability are established at 12, 16, and 18 years old, which play a vital role in determining the case's merits and verdict.

As such, our proposed bone age estimation system could aid in the accurate prediction of these threshold ages. The proposed study did, however, have certain shortcomings. Girls typically enter adulthood sooner than boys, generally between the ages of 8 and 13. Also, several dietary and hormonal factors influence bone growth metrics in healthy boys and girls at various pubertal phases. Yet, the given framework did not take gender roles into account while predicting age. The further drawback of this proposed work is the diminutive sample size, which could impede the effectiveness of deep learning algorithms.

5. Conclusion

This research aims to enhance the ability of deep learning networks for age prediction tasks. To this end, we offered a two-stage deep learning framework that combined a bone extraction model with an age evaluation model. The experiment revealed that transferring knowledge from a small, well-trained CNN classifier into the deep CNN model improved multi-classification performance over the existing deep learning models. Besides, the K-fold cross-validation technique assessed the model using various subsets of sample data while validating its structure for resilience. Moreover, the iterative validation approach utilized the knowledge from previous iterations, enhancing findings and accelerating the model's convergence in each cycle. Also, the experiment proved that the pre-trained weights from the final iteration aided in training deep-learning models without affecting the quality of the results. While the models showed varying performance levels, stacking utilized the strengths of each model to generate reliable age predictions. In future studies, it is necessary to expand the population size and the age group of the sample beyond 19 years old. Furthermore, the future framework should incorporate various biological factors other than radiographs for widespread application.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] M. Miranker, "A Comparison of Different Age Estimation Methods of the Adult Pelvis," *Journal of Forensic Sciences*, vol. 61, no. 5, pp. 1173–1179, Jun. 2016.

[2] B. Gümüş, E. Karavaş and O. Taydaş, "Can forensic radiological skeletal age estimation be performed by examining ischiopubic-ilioischial-iliopubic synchondrosis in computed tomography images?" *PLoS ONE*, vol. 17, no. 4, p. e0266682, Apr. 2022.

[3] E. Stan, A. Enache, C.O. Muresan, V. Ciocan, S. Ungureanu, A.C. Motofelea, A. Voicu and D. Costachescu, "Age Estimation through Hounsfield Unit Analysis of Pelvic Bone in the Romanian Population," *Diagnostics*, vol. 14, no. 18, pp. 2103–2103, Sep. 2024.

[4] S.R. Vempalli, V. Meshram, R.S. Shekhawat, B. Sureka, R. Shedge, M.A. NJ, P. Setia and T. Kanchan, "3D CT based age estimation from the pubic symphyseal surface in an Indian population using the Chen et al. method," *Anthropologischer Anzeiger*, vol. 81, no. 3, pp. 315–325, Jun. 2024.

[5] V. Warrier, R. Shedge, P.K. Garg, S.G. Dixit, K. Krishan and T. Kanchan, "Machine learning and regression analysis for age estimation from the iliac crest based on computed tomographic explorations in an Indian population," *Medicine, Science and the Law/Medicine, science and the law*, vol. 64, no. 3, pp. 204–216, Sep. 2023.

[6] F. Fan, X. Dong, X. Wu, R. Li, X. Dai, K. Zhang, F. Huang and Z. Deng, "An Evaluation of Statistical Models for Age Estimation and the Assessment of the 18-year Threshold using Conventional Pelvic Radiographs," *Forensic Science International*, vol. 314, pp. 110350–110350, Jun. 2020.

[7] K. Zhang, X. Dong, F. Fan and Z. Deng, "Age estimation based on pelvic ossification using regression models from conventional radiography," *International Journal of Legal Medicine*, vol. 130, no. 4, pp. 1143–1148, May 2016.

[8] Y. Li, Z. Huang, X. Dong, W. Liang, H. Xue, L. Zhang, Y. Zhang and Z. Deng, "Forensic age estimation for pelvic X-ray images using deep learning," *European Radiology*, vol. 29, no. 5, pp. 2322–2329, Nov. 2018.

[9] R. Akhade, A. Dhanorkar, J. Chawhan and J. Khanapuri, "Bone Age Estimation System Using Deep Learning," in *Proc. of the 5th International Conference on Advances in Science and Technology (ICAST)*, IEEE, 2022, pp. 1-5.

[10] L.Q. Peng, Y.C. Guo, L. Wan, T.A Liu, P. Wang, H. Zhao and Y.H. Wang, "Forensic bone age estimation of adolescent pelvis X-rays based on two-stage convolutional neural network," *International Journal of Legal Medicine*, vol. 136, no. 3, pp. 797–810, Jan. 2022.

[11] T. Bradshaw, Z. Huemann, J. Hu and A. Rahmim, "A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging," *Radiology*, vol. 5, no. 4, p. e220232, Jul. 2023.

[12] S. Prusty, S. Patnaik and S. K. Dash, "SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer," *Frontiers in Nanotechnology*, vol. 4, p. 972421, Aug. 2022.

[13] T.R. Mahesh, O. Geman, M. Margala, and M. Guduri, "The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification," *Healthcare Analytics*, vol. 4, p. 100247, Dec. 2023.

[14] J. Kaliappan, A.R. Bagepalli, S. Almal, R. Mishra, Y.-C. Hu and K. Srinivasan, "Impact of Cross-Validation on Machine Learning Models for Early Detection of Intrauterine Fetal Demise," *Diagnostics*, vol. 13, no. 10, pp. 1692–1692, May 2023.

[15] M.N. Raihen and S. Akter, "Comparative Assessment of Several Effective Machine Learning Classification Methods for Maternal Health Risk," *Computational Journal of Mathematical and Statistical Sciences*, vol. 3, no. 1, pp. 161–176, Apr. 2024.

[16] D. Rastogi, P. Johri, V. Tiwari and A. A. Elngar, "Multi-class classification of brain tumour magnetic resonance images using multi-branch network with inception block and five-fold cross validation deep learning framework," *Biomedical Signal Processing and Control*, vol. 88, p. 105602, Feb. 2024.

[17] N. Nivedita and S. Solanki, "Enhancing the accuracy of automatic

bone age estimation using optimized CNN model on X-Ray images," in *Communications in computer and information science*, 2024, pp. 329–340.

[18] A. Davila, J. Colan and Y. Hasegawa, "Comparison of fine-tuning strategies for transfer learning in medical image classification," *Image and vision computing*, vol. 146, pp. 105012–105012, Apr. 2024.

[19] O. Ronneberger, P. Fischer and T. Brox, "U-NET: Convolutional Networks for Biomedical Image Segmentation," in *Lecture notes in computer science*, 2015, pp. 234–241.

[20] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz and B. Glocker, "Attention u-net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999.*