Appendix 1

Ablation study

The aim of ablation study is to justify the right choice for the two most important hyper-parameters contributing to the model performance: image resolution and the inclusion of the co-variate data. The best image resolution is found among square-scaled (1:1) of 128, 256, 320, 480, 512, and 1,024.

To speed up the ablation training sessions, only the 3rd fold of the cross-validation splits introduced in the dataset section of the paper is used to validate the model, assuming that the rest of the folds generalize to the same outcome. As a matter of fact, the results presented in this section, and those in the results section of the paper must not be compared.

Similar to the post-processing steps explained in the previous section, all the predicted BMDs are converted from Lunar to Hologic, and T-scores are derived from the calibrated BMDs using female peak bone mass from NHANES III (18). The AUROC metric is used to evaluate the trained models, and the impact of each hyper-parameter on the model performance is evaluated independent of one another, to only highlight that hyper-parameter. Finally, an iterative feature ablation process is done on the BaseDT model to demonstrate the contribution of every individual bone features on the model performance.

Choice of the image backbone architecture and the training strategy details

We treated the choice of the backbones and the fusion module architecture as additional hyper-parameters, and performed an extensive search, by using the validation split, to tune them to the best setting. Specifically, we limited our image backbone search to the commonly used architectures in the computer vision domain (with rich literature in medical applications) such as ResNet (28), EfficientNet (29), and InceptionV3 (30). In particular, InceptionV3 has proven its success in one of our previous research papers in (31). Moreover, we tested various MLP architectures for the co-variate data and the fusion module by trying different fully-connected layer sizes, activation functions, and the dropout probability.

We used transfer learning, i.e., the network weights were initialized with pretrained weights on the ImageNet dataset, mainly due to lack of sufficient data to train a model from scratch. As a common practice in transfer learning, the weights of the shallower blocks of the model were frozen, and the last 3 inception blocks of 7a, 7b, and 7c were fine-tuned using the OAI dataset. The initial learning rate was set to 1e-6, and it was scheduled to decrease by half at epoch 100. We trained the entire model for 500 epochs, while employing an early stopping technique that would terminate the training process if there was no significant decay in the validation loss for 20 consecutive epochs. This ultimately helped the model to generalize better to the test set samples.

Contribution of the image resolution and the co-variate data

Different image sizes are experimented by using only the X-ray images (not including the co-variate data). Then 6 models (for each image size) are trained, and their AUROC performance on the test dataset is evaluated. As it is depicted in *Figure S1*, both the pelvis and the knee models AUROC increases with increasing image size up to around 480 to 512, then it sharply declines at 1,024, which is due to the overfitting problem. However, the hand model shows a different trend with the image size 256 being the best one from where the AUROC starts to decline. One of the main reasons for this behavior can be attributed to the size of the hand X-rays dataset, which is smaller than the knee dataset, and almost the same as the pelvis dataset. Now the pelvis model trend is very similar to the knee model trend, although it has a similar number of samples to the hand dataset, it has learned the decisive features to predict the femoral neck BMD quite well, mainly because the femoral neck is in fact present in the X-ray image.

It was previously shown in the paper that the X-ray images play an important role in improving the model performance. Here, the contribution of the co-variate data (sex and age) to the performance of the model is studied by training knee, hand, and pelvis models with and without the co-variate data and comparing their AUROC. *Table S1* summarizes the results. Interestingly, the performance of all the body parts models improves when the co-variate data are used.

https://dx.doi.org/10.21037/jmai-23-10

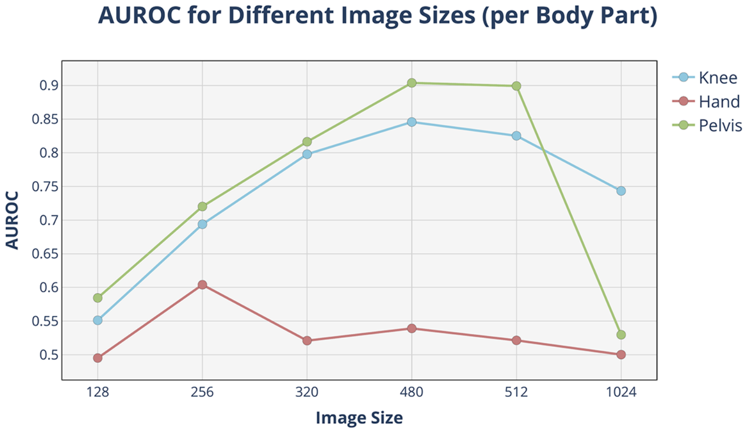

Figure S1 The varying AUROC performance of the DL model with different image sizes. AUROC, area under the receiver operating characteristic curve; DL, deep learning.

Table S1 The AUROC performance of the knee, hand, and pelvis models with and without using the co-variate data

Body part	Input source	AUROC
Hand	X-ray + co-variate data	0.8294
	X-ray	0.7907
Knee	X-ray + co-variate data	0.8765
	X-ray	0.8537
Pelvis	X-ray + co-variate data	0.9351
	X-ray	0.9296

AUROC, area under the receiver operating characteristic curve.

Extra co-variate data contribution to the BaseDT model

First, we show the correlation within the co-variate data features, and between the co-variate data features and the target feature, to find any high (positive or negative) correlation between any specific feature, and the target label. *Figure S2* displays the heatmap calculated using the Pearson correlation coefficient (32). The correlation coefficients are between −1 (maximum negative correlation) and 1 (maximum positive correlation). Clearly, there is maximum positive correlation between interview_age (age in months) and ageyears (age in years) features, which is expected. Moreover, trth (trabecular thickness) and trn (trabecular number) show very high correlation between one another (0.79), and with bvf (bone volume fraction) too (0.93 with trth and 0.95 with trn). Most importantly, trn (trabecular number) and ageyears (age in years) show relatively high correlation (positive and negative, respectively) with the target feature, neckbmd (femoral neck BMD), which are definitely helpful as input features for any model to generate a more accurate prediction.

It is shown in *Table S2* how these additional features cumulatively contributed to the performance of the BaseDT model for hand, on the female group. The BaseDT model was trained on varying subsets of features using 5-fold cross-validation on the same splits used to train the main model. As it is shown in *Table S2*, as the features are dropped in a random order from the list of features, the model sees a declining trend in accuracy, sensitivity, and specificity. In fact, the trabecular thickness (tth) and the trabecular number (tn) co-variate data, which are shown to boost up the overall performance of the BaseDT model significantly, are not included as additional co-variate data next to the X-ray images in the DL model.

https://dx.doi.org/10.21037/jmai-23-10

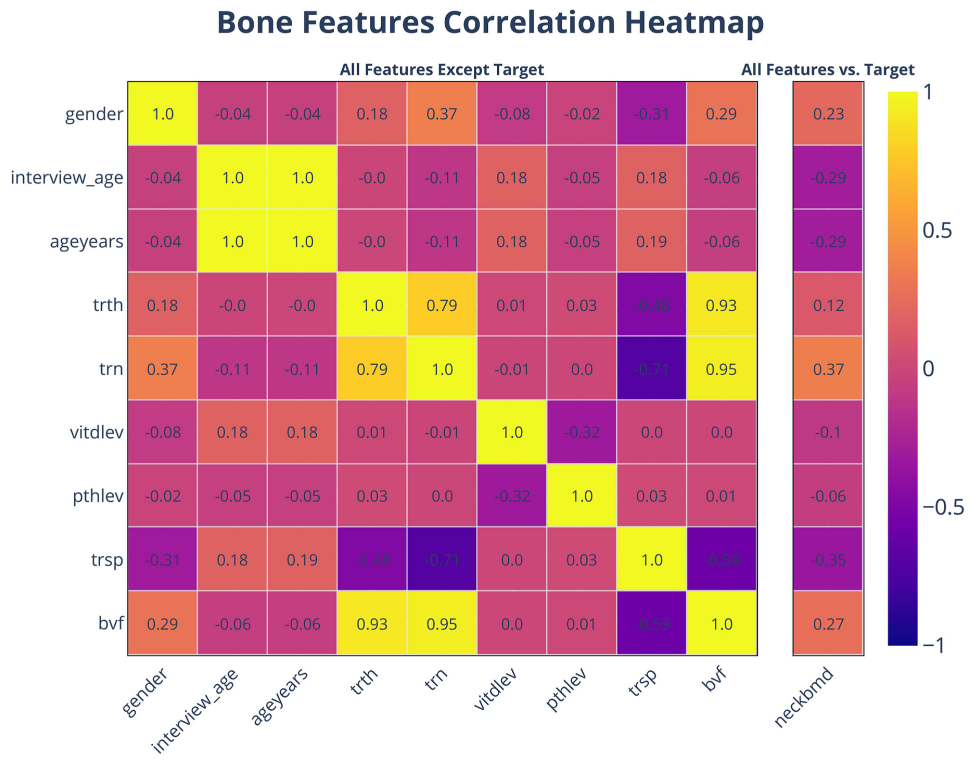

Figure S2 The correlation heatmap within the co-variate data features (left), and between the co-variate data features and the target feature (right) using the Pearson correlation coefficient. interview_age, age in months; ageyears, age in years; trth, trabecular thickness; trn, trabecular number; vitdlev, vitamin D level; pthlev, intact parathyroid level; trsp, trabecular spacing; bvf, bone volume fraction; neckbmd, femoral neck bone mineral density.

Table S2 Declining performance of the BaseDT model as the number of co-variate data features decrease

Features	Accuracy	Sensitivity	Specificity
g + ia + ay + hs + tth + tn + vl + pl + ts + bvf	0.80	0.7806	0.81
g + ia + ay + hs + tth + tn + vl + pl + ts	0.78	0.7371	0.81
g + ia + ay + hs + tth + tn + vl + pl	0.76	0.7371	0.77
g + ia + ay + hs + tth + tn + vl	0.74	0.6936	0.77
g + ia + ay + hs + tth + tn	0.74	0.6936	0.77
g + ia + ay + hs + tth	0.72	0.6936	0.73
g + ia + ay + hs	0.66	0.6067	0.69
g + ia + ay	0.68	0.6001	0.69

BaseDT, baseline decision tree; g, gender; ia, age in months; ay, age in years; hs, hipside; tth, trabecular thickness; tn, trabecular number; vl, vitamin D level; pl, intact pth level; ts, trabecular spacing; bvf, bone volume fraction.

References

28. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-8.
29. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, 2019:6105-14.
30. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:2818-26.
31. Halabi SS, Prevedello LM, Kalpathy-Cramer J, et al. The RSNA Pediatric Bone Age Machine Learning Challenge. Radiology 2019;290:498-503.
32. Benesty J, Chen J, Huang Y, et al. Pearson Correlation Coefficient. In: Cohen I, Huang Y, Chen J, et al. Noise Reduction in Speech Processing. Berlin: Springer, 2009:1-4.