Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Improving Automated Pediatric Bone Age Estimation Using Ensembles of Models from the 2017 RSNA Machine Learning Challenge

Ian Pan, MA • Hans Henrik Thodberg, PhD • Safwan S. Halabi, MD • Jayashree Kalpathy-Cramer, PhD • David B. Larson, MD, MBA

From the Department of Radiology, Warren Alpert Medical School, Brown University, 593 Eddy St, Providence, RI 02903 (I.P.); Department of Diagnostic Imaging, Rhode Island Hospital, Providence, RI (I.P.); Visiana, Hørsholm, Denmark (H.H.T.); Department of Radiology, Stanford University, Palo Alto, Calif (S.S.H., D.B.L.); and Department of Radiology, Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Boston, Mass (J.K.C.). Received April 16, 2019; revision requested April 17; revision received July 12; accepted August 23. **Address correspondence to** I.P. (e-mail: *ianpan358@gmail.com*).

See also the commentary by Siegel in this issue.

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2019; 1(6):e190053 • https://doi.org/10.1148/ryai.2019190053 • Content codes: IN MK PD

Purpose: To investigate improvements in performance for automatic bone age estimation that can be gained through model ensembling.

Materials and Methods: A total of 48 submissions from the 2017 RSNA Pediatric Bone Age Machine Learning Challenge were used. Participants were provided with 12 611 pediatric hand radiographs with bone ages determined by a pediatric radiologist to develop models for bone age determination. The final results were determined using a test set of 200 radiographs labeled with the weighted average of six ratings. The mean pairwise model correlation and performance of all possible model combinations for ensembles of up to 10 models using the mean absolute deviation (MAD) were evaluated. A bootstrap analysis using the 200 test radiographs was conducted to estimate the true generalization MAD.

Results: The estimated generalization MAD of a single model was 4.55 months. The best-performing ensemble consisted of four models with an MAD of 3.79 months. The mean pairwise correlation of models within this ensemble was 0.47. In comparison, the lowest achievable MAD by combining the highest-ranking models based on individual scores was 3.93 months using eight models with a mean pairwise model correlation of 0.67.

Conclusion: Combining less-correlated, high-performing models resulted in better performance than naively combining the top-performing models. Machine learning competitions within radiology should be encouraged to spur development of heterogeneous models whose predictions can be combined to achieve optimal performance.

Supplemental material is available for this article.

Ensemble learning is a method in machine learning in which different models designed to accomplish the same task are combined into a single model (1–3). This method has been used in a number of computer vision and machine learning models published in the radiology literature (4–6).

Model heterogeneity is an important aspect of ensemble learning. Ensembles tend to perform best when each of the individual models performs well in its own right and the correlation among individual model predictions is relatively low (7). Achieving such diversity can be facilitated by developing a number of models using a variety of techniques and selecting the most accurate combination. Because ensembles benefit from low correlation between model predictions, the greater the underlying differences in approach, the greater the improvement, as long as they achieve similar performance. In this respect, a competition in which participants are encouraged to submit their best models provides an ideal setting from which to ensemble high-performing models that use different techniques.

The 2017 RSNA Pediatric Bone Age Machine Learning Challenge (8) was a competition in which participants were provided with a set of hand radiographs for determination of bone age, along with estimated bone ages according to the Greulich and Pyle atlas (9), to be used as training data. The challenge provided a unique opportunity to test the power of ensembling the 48 submitted computer vision models.

The purpose of this article was to investigate improvements in performance for automatic bone age estimation that can be gained through ensembling heterogeneous models from the 2017 RSNA Pediatric Bone Age Machine Learning Challenge.

Materials and Methods

The institutional review boards at Stanford University School of Medicine and University of Colorado Health Sciences Center approved the curation and use of pediatric hand radiographs for the purposes of this challenge. Patient consent was waived after approval by the institutional review board.

The 2017 RSNA Pediatric Bone Age Machine Learning Challenge was an open competition sponsored by the

Abbreviations

AMC = all model combinations, MAD = mean absolute deviation, meAD = median absolute deviation

Summary

Ensembles created using models submitted to an open competition convincingly outperformed single-model prediction for bone age assessment.

Key Points

- Model ensembles were able to decrease the generalization error of bone age prediction from a mean absolute deviation of 4.55 months (one model) to 3.79 months (four models).
- Combining less-correlated and relatively high-performing models resulted in better ensembles than naively combining the top-performing individual models.

RSNA and administered by the RSNA's Radiology Informatics Committee and other volunteers. A full description of the challenge was described by Halabi et al (8).

For the challenge, participants were provided with training and validation sets of 12 611 and 1425 annotated hand radiographs, respectively, to develop automated bone age prediction models over 10 weeks. A test set of 200 images (100 in male subjects, 100 in female subjects) was subsequently released; participants submitted bone age estimates on the test set to determine the final standings.

Images and annotations were donated to the RSNA by the authors of a previously published study of a deep learning model (10). All images and annotations were obtained from the clinical picture archiving and communication system from Lucile Packard Children's Hospital at Stanford and Children's Hospital Colorado.

Ground-truth values for the test set were calculated as a weighted average of six reviewers' assessments, based on each reviewers' performance relative to the other reviewers. The mean reviewer mean absolute deviation (MAD) relative to the ground-truth estimate was 5.8 months. The relative weightings for the reviewers ranged from a minimum of 0.14 to a maximum of 0.21, with a mean of 0.17.

After calculating the MAD for each submission, the identities of the submitting individuals and institutions were removed for the purpose of this analysis. Each model was assigned an identifier number, in ascending order of MAD. Each submission consisted of model-based predictions of the bone age in months for each of the 200 test images. Ensembles were formed by calculating the unweighted mean of the predictions for each model in the ensemble.

Statistical Analysis

Statistical analysis for this study was conducted primarily through simulation using the NumPy scientific computing library in the Python 2.7 programming language (Python Software Foundation, Wilmington, Del; *https://www.python.org*). We investigated ensembles of up to 10 models. All models were equally weighted in each ensemble. To calculate performance of various ensembles, we randomly divided the challenge test set into 1000 simulated validation-test splits. This prevented the use of the same data for both model selection and model evaluation and allowed us to more accurately assess single-model performance.

Figure 1 illustrates our experimental design. Each simulated validation-test split represents an experiment. The data were split into a validation set for ensemble selection (50%, 100 images) and a test set for ensemble evaluation (50%,

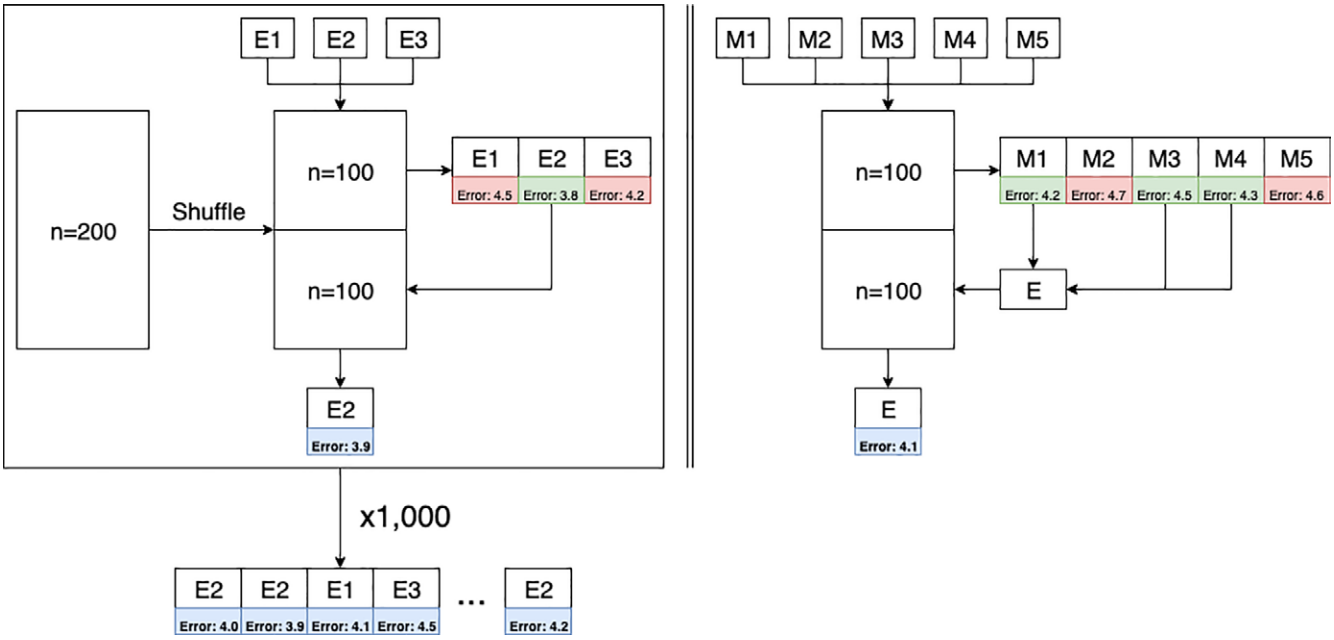

Figure 1: Schematic illustrates the experimental design. E and M refer to theoretical ensembles and individual models, respectively. Left: Ensemble selection and evaluation process of all-model-combinations method where there exist three possible ensembles. The original dataset is randomly split into 50% validation and 50% test. The validation set is used to determine the best ensemble, and its performance is determined on the test set. Right: Ensemble selection process of top-N method with a pool of five models forming three-model ensembles. The evaluation process remains the same.

100 images). We also performed a sensitivity analysis using the median absolute deviation (meAD) as an alternative evaluation metric. To select the best ensemble, we first calculated the MAD of all possible ensembles of a given size (eg, three models) on the simulated validation set and selected the ensemble with the lowest MAD to predict on the simulated test set, which we will refer to as the all-model-combinations (AMC) method. This process was repeated 1000 times, and the average over the 1000 simulated test set MADs was taken as an approximation of the true generalization error of the ensemble. Using the distribution obtained from the 1000 experiments, 95% confidence intervals were calculated for the ensemble MAD values. Each of the 1000 experiments involved an ensemble selection phase on a slightly different validation set of 100 images; thus, it is possible that the best ensemble contains different models across the 1000 experiments. For comparison, we also created ensembles by simply combining the top two to 10 models by individual MAD on the simulated validation set, referred to as the top-N method, instead of searching over the space of all possible ensembles. To examine the effect of ensembling on outliers (ie, cases with large prediction errors), we calculated the MADs ≥ 95th percentile (MAD_{95}) by taking the mean over absolute deviations at and above the 95th percentile in the test set. To determine potential complementarity of models, we calculated the pairwise Pearson correlations of the residuals of the estimates of all model pairs (R_p). We estimated the MAD of two-model ensembles using the Equation below, which demonstrates the trade-off between model correlation and model performance (11):

$$\mathrm{MAD}_{\mathrm{ensemble}} = \frac{1}{2}(\mathrm{MAD}_1 + \mathrm{MAD}_2)\sqrt{\frac{1+R_P}{2}}.$$

We defined the mean model correlation for a model (R_{MM}) over a given model space to be the average of its pairwise correlations with every other model in that space (ie, average

Table 1: Summary Statistics of Individual Models

Parameter	MAD (mo)	RMSD (mo)	R_{MM}	$R_{MM,<6}$
Minimum	4.27	5.64	0.30	0.41
Maximum	34.16	44.53	0.61	0.70
Mean	8.64	11.28	0.48	0.62
Median	5.99	7.91	0.50	0.63

Note.—Data are mean absolute difference (MAD) in months, root-mean-square deviation (RMSD) in months, mean model correlations over all 48 models (R_{MM}), and mean model correlations over 24 models with MAD less than 6 months ($R_{MM,<6}$).

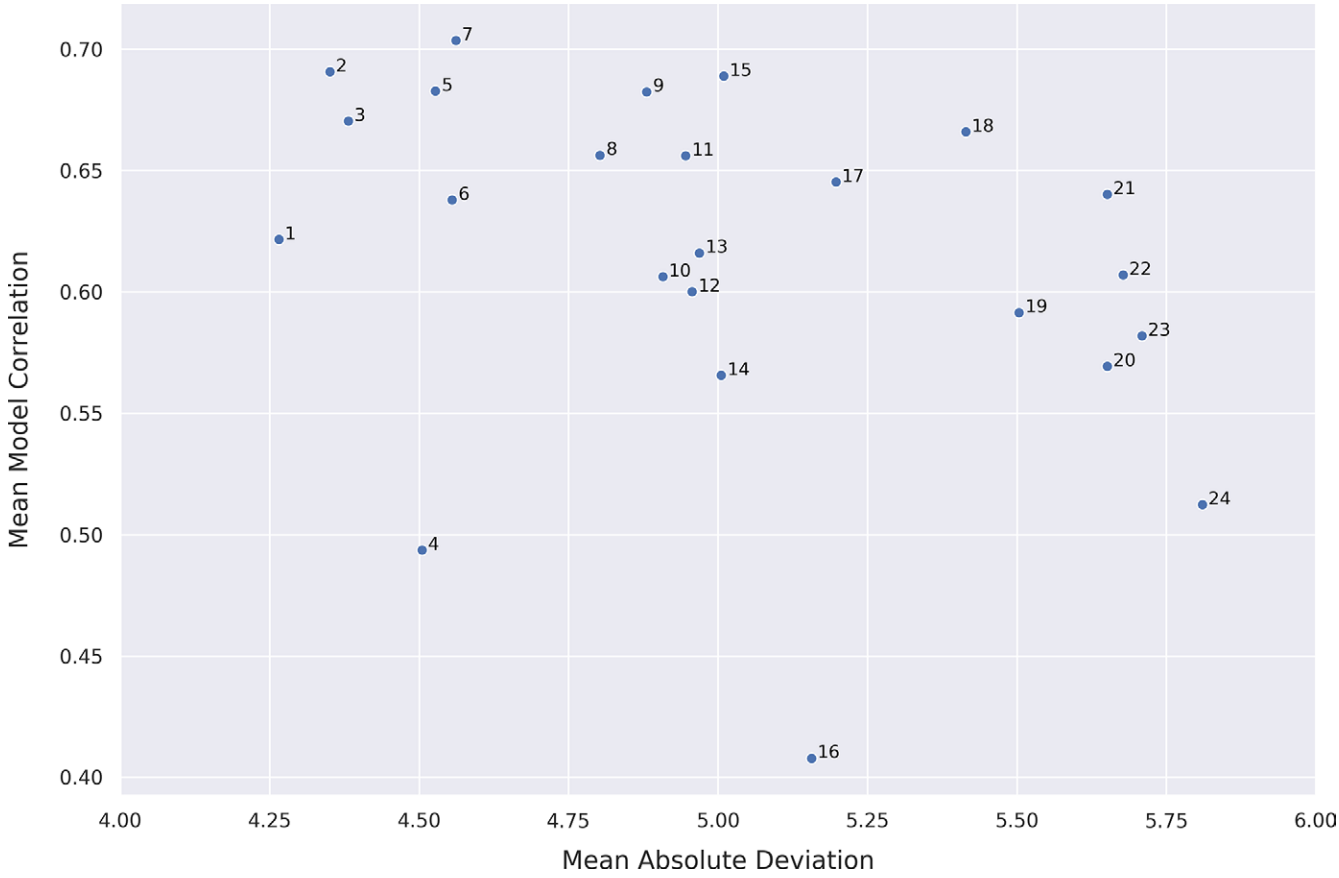

Figure 2: Mean model correlation compared with mean absolute difference (MAD) in months. Mean model correlation that was averaged over the 24 models with an MAD of less than 6 months is shown here. Each point is labeled with the corresponding model number. Models toward the bottom of the graph have low correlations with other models, suggesting that they would be beneficial in an ensemble.

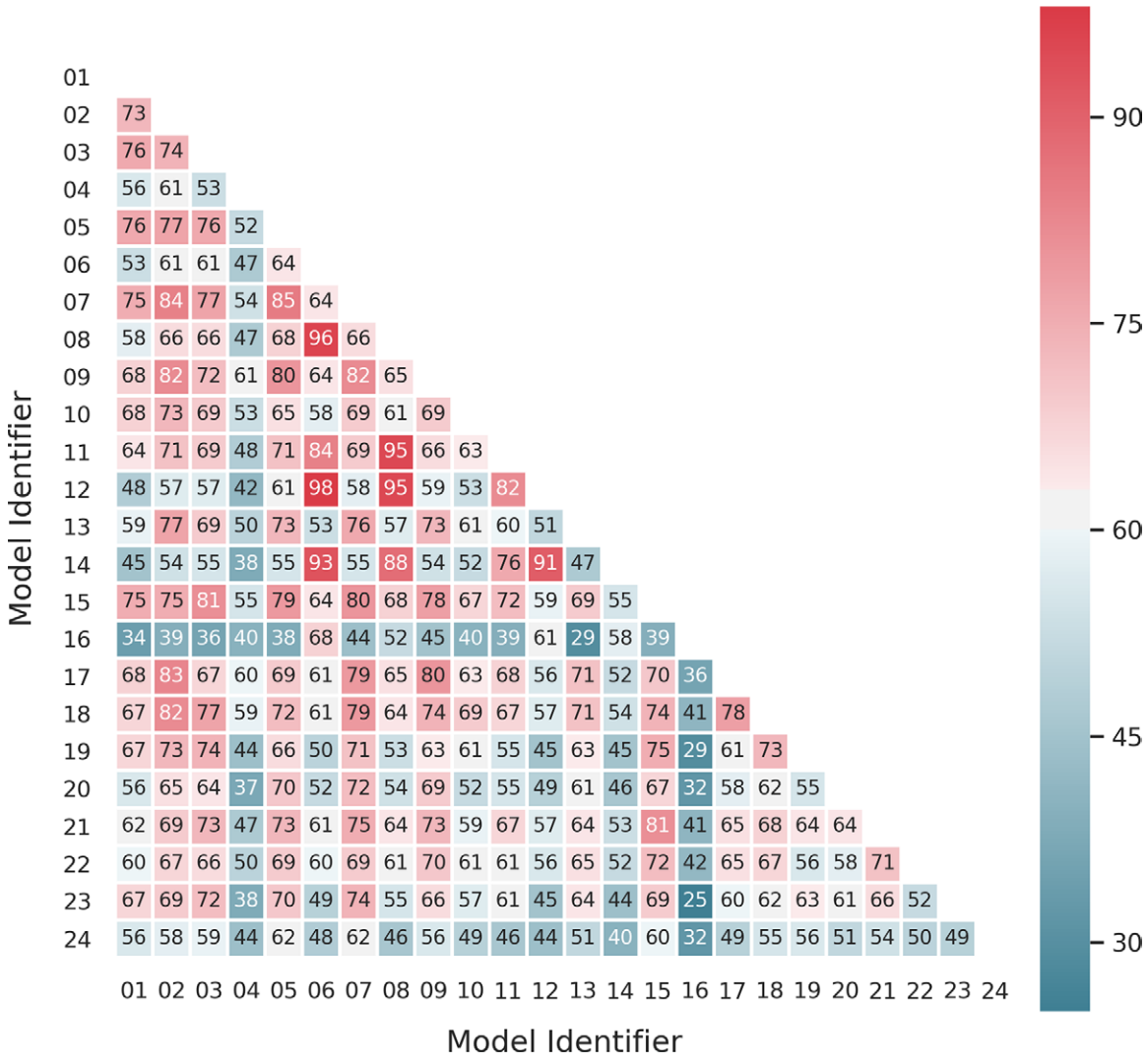

Figure 3: Paired model correlations for pairs of models with mean absolute deviation less than 6 months. Models 4 and 16 stand out as models with low correlation (blue) with other models. Red indicates higher model correlations.

of all possible R_p). We referred to $R_{MM,<6}$ as the mean model correlation over all models with MAD less than 6 months and $R_{MM,ensemble}$ as the mean model correlation over all models within a given ensemble.

Results

Descriptive statistics for the training and test sets were previously reported by Larson et al (10). Performance of the 48 models submitted for the challenge were previously reported by Halabi et al (8). Individual model performance, as measured by MAD, ranged from 4.27 to 34.16 months with a median of 5.99 months.

The R_{MM} over all 48 models ranged from 0.30 to 0.61 with a median of 0.50; when excluding models with MAD greater than 6 months, the $R_{MM,<6}$ over the resulting 24 models ranged from 0.41 to 0.70 with a median of 0.63. Summary statistics for the performance of the individual models are shown in Table 1, and a graph of MAD versus $R_{MM,<6}$ is shown in Figure 2. Individual MAD, R_{MM}, and $R_{MM,<6}$ values for each submitted model are available in Table E1 (supplement). We observed that pairs consisting of models with higher individual MADs did not necessarily outperform other model pairs. Figure 3 is a graphical depiction of the R_p across all pairs of models with an MAD of less than 6 months. A corresponding depiction of the two-model ensemble MAD, calculated over all 200 test cases, is shown in Figure 4. The top 10 two-model ensembles are shown in Table 2 with the actual and estimated MADs using the Equation. The mean absolute percentage difference between actual and estimated MADs was 3.8%. The mean absolute percentage difference was 1.8% for ensembles with MAD less than 6 months, 3.9% for ensembles with MAD between 6 and 10 months, and 8.5% for ensembles with MAD greater than 10 months.

Two models emerged as potentially the most complementary to other models, with relatively low MAD and R_{MM}: models 4 and 16. Model 4 had an MAD and $R_{MM,<6}$ of 4.50 months and 0.49, respectively, and model 16 had an MAD and $R_{MM,<6}$ of 5.16 months and 0.41, respectively. This is evident in Figure 2, as these models stand out with low correlation with other models.

Upon investigation, model 4 utilized nondeep machine learning (a modular pipeline of active appearance models, principal component analysis, and linear regression), which differed from the other top models, which used deep learning and convolutional neural networks. We did not receive a methods description for model 16 because the model contributors did

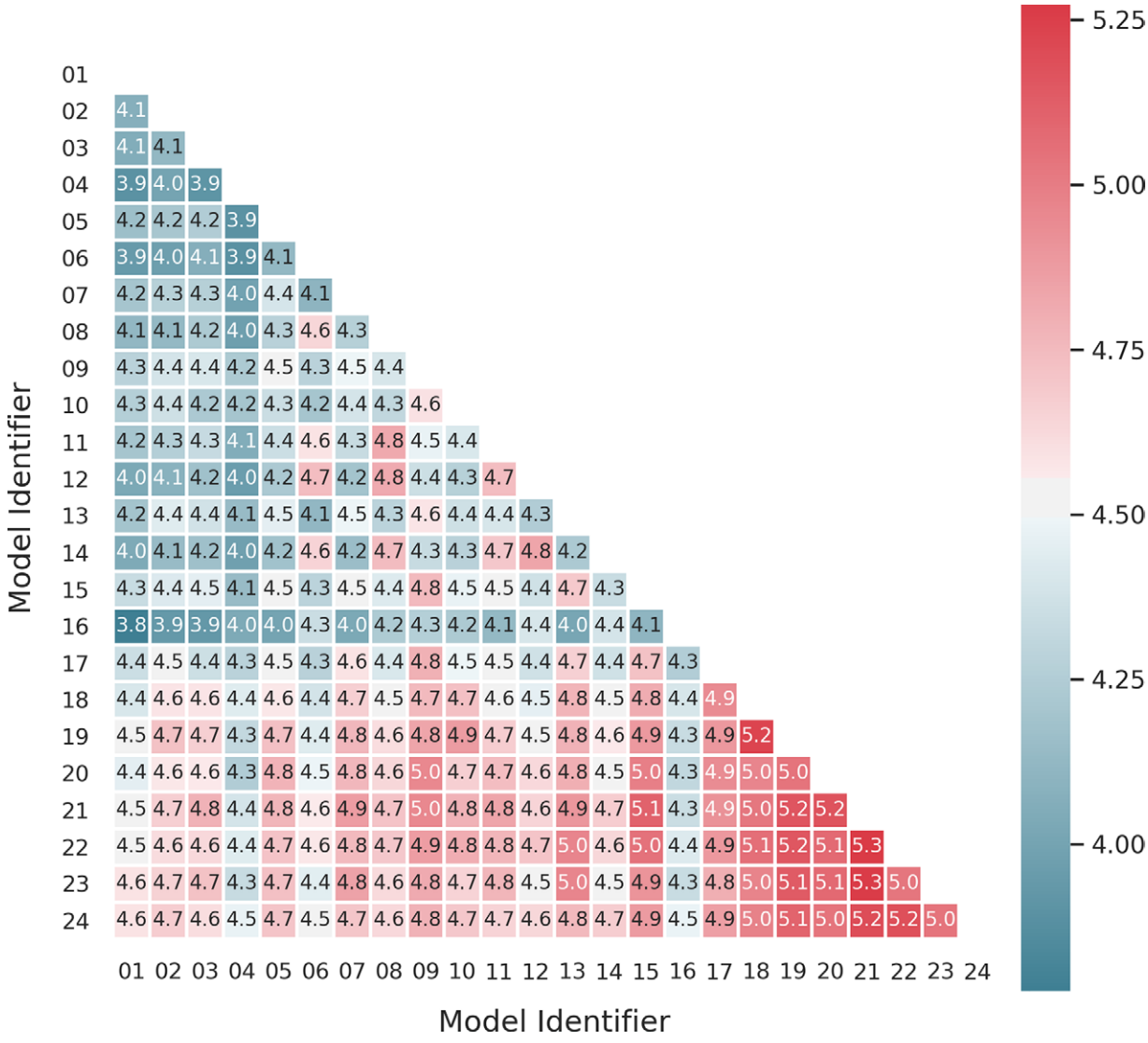

Figure 4: The two-model ensemble mean absolute deviations (MADs) for pairs of models with MAD less than 6 months. MADs are calculated over all 200 cases in the original challenge test set. Red (worse) and blue (better) indicate higher and lower MADs, respectively.

not respond to multiple e-mail requests for a description.

The performance of ensembles of different sizes based on MAD is presented in Table 3. Results of the sensitivity analysis using meAD can be found in Table E2 (supplement). The estimated generalization MAD of a single model was 4.55 months. The minimum MAD was 3.79 months, achieved with four models using AMC. The $R_{MM,ensemble}$ over these four models was 0.47. In comparison, the estimated MAD of a four-model ensemble created using top-N was 4.01 months with an average R_{MM} of 0.68. The highest performing top-N ensemble achieved an MAD of 3.93 months using eight models with an $R_{MM,ensemble}$ of 0.67. Evaluating the meAD demonstrated a similar trend: a single

Table 2: Top 10 Two-Model Ensembles

Model Ensemble	Actual MAD (mo)	Estimated MAD (mo)	Mean Model MAD (mo)	R_p
1 and 16	3.78	3.85	4.71	0.34
4 and 6	3.89	3.88	4.53	0.47
1 and 4	3.90	3.87	4.38	0.56
4 and 5	3.90	3.94	4.52	0.52
3 and 4	3.91	3.88	4.44	0.53
3 and 16	3.93	3.93	4.77	0.36
2 and 16	3.94	3.96	4.75	0.39
1 and 6	3.94	3.86	4.41	0.53
4 and 12	3.96	3.98	4.73	0.42
4 and 8	3.97	3.99	4.65	0.47

Note.—Data are as determined by mean absolute deviation (MAD) in months calculated over all 200 cases. Estimated MAD is based on the Equation. Mean of the individual model MADs and model correlations are shown as well.

Table 3: Ensemble Mean Absolute Difference according to Number of Models

No. of Models	Ensemble MAD (mo)				$R_{MM,ensemble}$		Mean Model MAD (mo)	
	AMC	MAD_{95}	Top-N	MAD_{95}	AMC	Top-N	AMC	Top-N
1	4.55 (3.99, 5.14)	15.07	4.55 (3.99, 5.14)	15.07			4.16	4.16
2	4.06 (3.50, 4.60)	13.54	4.20 (3.69, 4.73)	13.82	0.40	0.68	4.49	4.22
3	3.86 (3.35, 4.44)	12.74	4.07 (3.55, 4.61)	13.25	0.44	0.68	4.57	4.28
4	3.79 (3.30, 4.32)	12.54	4.01 (3.48, 4.56)	12.88	0.47	0.68	4.62	4.32
5	3.79 (3.31, 4.30)	12.40	3.97 (3.43, 4.52)	12.66	0.49	0.68	4.65	4.36
6	3.80 (3.32, 4.34)	12.40	3.95 (3.45, 4.50)	12.51	0.51	0.67	4.67	4.40
7	3.80 (3.34, 4.29)	12.37	3.94 (3.46, 4.48)	12.44	0.53	0.67	4.69	4.44
8	3.81 (3.34, 4.33)	12.42	3.93 (3.46, 4.46)	12.41	0.54	0.67	4.70	4.47
9	3.83 (3.35, 4.29)	12.48	3.94 (3.47, 4.47)	12.45	0.55	0.66	4.72	4.51
10	3.84 (3.37, 4.32)	12.51	3.93 (3.45, 4.46)	12.49	0.56	0.66	4.73	4.54

Note.—Two sets of ensembles are presented here: all possible ensembles over all model combinations (AMC) and top-N ensembles (top-N) based on individual model performance. Data are the mean and 95% confidence interval in parentheses over the 1000 experiments for the ensemble mean absolute deviation (MAD) in months and the means for the MAD_{95}, $R_{MM,ensemble}$, and average individual model MAD. $R_{MM,ensemble}$ values are calculated over all models included in the ensemble on the simulated training (model selection) split. The average individual model MADs (mean model MAD) were calculated by taking the average of the MADs on the validation split of each model in the ensemble.

Table 4: Models Comprising the Best All Model Combinations Ensembles in Table 2

	All Model Combinations		
No. of Models	MAD (mo)	Best Ensemble	Runner-up
1	4.55	1 (44%)	2 (22%)
2	4.06	1–16 (42%)	4–5 (10%)
3	3.86	1–4–16 (37%)	1–3–16 (10%)
4	3.79	1–3–4–16 (21%)	1–13–14–16 (10%)
5	3.79	1–4–13–14–16 (8%)	1–3–4–6–16 (5%)
6	3.80	1–3–4–13–14–16 (5%)	1–2–3–4–13–16 (3%)
7	3.80	1–3–4–10–13–14–16 (4%)	1–2–3–4–13–14–16 (3%)
8	3.81	1–2–3–4–10–13–14–16 (4%)	1–3–4–6–10–13–14–16 (3%)
9	3.83	1–2–3–4–6–10–13–14–16 (4%)	1–3–4–6–10–12–13–14–16 (4%)
10	3.84	1–2–3–4–6–10–12–13–14–16 (4%)	1–2–3–4–6–10–13–14–16–19 (3%)

Note.—Because we assessed model performance over 1000 simulated train-test splits, a different combination of models could have been selected in each iteration. We report the most commonly occurring model combination with the percentage of iterations producing that model combination. The four-model ensemble achieved the lowest mean absolute difference (MAD) in months.

model achieved a meAD of 3.60 months, and the best ensemble consisted of five models with a meAD of 3.14 months. However, AMC did not outperform top-N. The optimal top-N ensemble consisted of six models with a meAD of 3.11 months. The averages of the individual model MADs are by design consistently lower for top-N ensembles, although the $R_{MM,ensemble}$ and ensemble MADs are higher. For each ensemble, the top two model combinations that occurred most often throughout the 1000 simulated validation-test sets are shown in Table 4. The MAD_{95} for the best ensemble (AMC, four models) was 12.54 months. The seven-model AMC ensemble achieved the lowest MAD_{95} of 12.37 months, an improvement of 2.70 months over a single model.

Figure 5 presents head-to-head ensemble comparisons by examining the percentage of experiments in which one ensemble was superior to another. All ensembles convincingly outperformed a single model in 97%–100% of experiments. In addition, AMC outperforms top-N at a given model size over 80% of the time. The optimal AMC ensemble (four models) achieves the lowest MAD in 24% of experiments, the highest among all ensembles. Collectively, AMC ensembles containing three to six models achieved the lowest MAD in 71% of experiments, whereas top-N ensembles achieved the lowest MAD in only 9% of experiments.

Discussion

By forming ensembles of machine learning models, we improved on the generalization error of a single model from an MAD of 4.55 months to 3.79 months. In comparison, the average MAD for the human reviewers was 5.8 months. We also note that the estimated generalization error for a single model was greater than the 4.26 months achieved by the challenge winner.

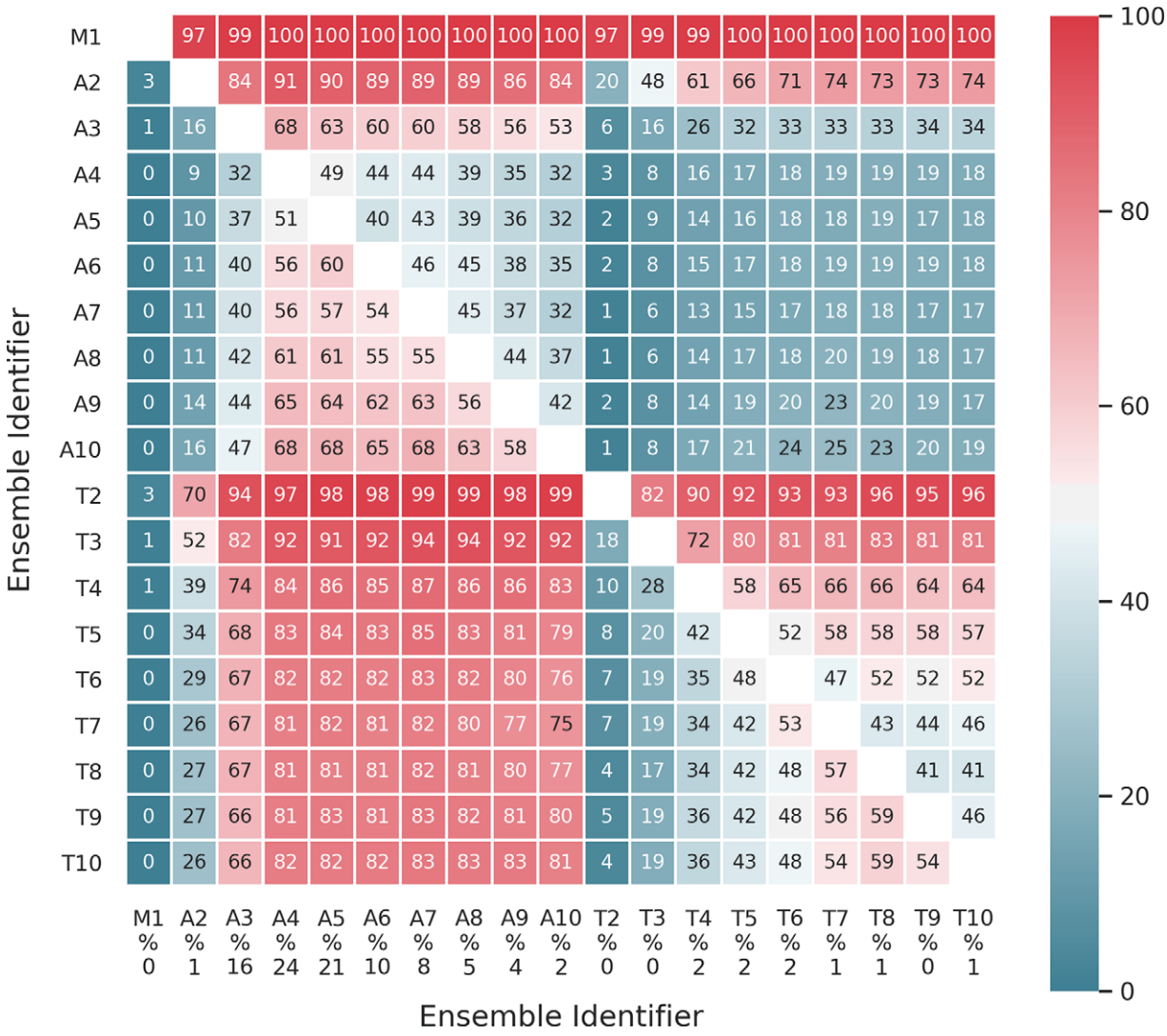

Figure 5: Head-to-head ensemble comparisons across 1000 experiments. The number displayed is the percentage of experiments where the x-axis ensemble outperformed the y-axis ensemble. M1 represents a single model. A and T refer to all model combinations and top-N ensembles, respectively, and the appended number refers to the number of models in the ensemble. Under each x-axis ensemble is the percentage of experiments where that ensemble achieved the lowest mean absolute deviation.

As hypothesized, we found that some models were more complementary than others in that when they were combined with other models, they brought about a relatively large performance increase. We also demonstrated that the best ensembles do not necessarily arise from combining the top-performing models. Upon further investigation, we found that the methodology used for one of the highly complementary models (model 4) substantially differed from that of the other top-performing models in that it used nondeep rather than deep learning. The developer of the other highly complementary model (model 16) did not respond to multiple requests for a description of the model; we suspect that their method also substantially differed from the other models.

While combining less-correlated models resulted in better ensembles when evaluating MAD, the same effect was not observed when evaluating meAD. Although ensembling still resulted in performance improvements, we found that slightly better results were achieved by simply combining the top-performing individual models versus searching over all possible ensembles. This suggests that model correlation is more relevant for ensemble formation for means-based versus median-based metrics. We advocate for the use of means-based metrics in medical applications, as medians underweight outliers, which arguably hold more importance from a risk analysis perspective.

We found that most of the benefit of ensembling was exhausted after combining just three models. While most of the benefit of ensembling was realized by the combination of a few models, it is important to note that this analysis was greatly facilitated by the contribution of many models in the competition. Without the development and testing of numerous models, it would not have been possible to know which two models should be combined and how they would perform relative to other models.

Dietterich described three fundamental sources of error in machine learning that ensembles can help overcome: statistical, computational, and representational (7). Statistical

error refers to the fact that, with limited training data, learning algorithms can find many different imperfect hypotheses that all give the same accuracy on the training data. The ensemble of models overcomes this error by averaging the votes of the different models. Computational error refers to the fact that learning algorithms may get stuck in local optima. An ensemble developed from models constructed by starting from different initializations may provide a more accurate approximation of the unknown function than any of the individual models. Representational error refers to the limitations of machine learning methods that either are limited in the number of hypotheses that can be used, or that can be generated based on a finite training sample. Combining the results of several models can essentially expand the number of hypotheses that can be considered or overcome limitations of model overfitting.

Our results call attention to a concept that has substantial practical implications as computer vision and other machine learning algorithms begin to move from research to the clinical environment, namely that the best results are likely to be achieved by combining multiple accurate and diverse models rather than from single models alone. Thus, practitioners aiming to incorporate machine learning algorithms into their workflow would benefit from having predictions obtained from different models, similar to how the accuracy of a radiologic interpretation can be bolstered with multiple readers. Human readers can be considered as unique models in their own right, and ultimately, we believe the most robust solutions will combine automated models with human assessment (12–15).

Our results also highlight the importance of open competitions, such as the 2017 RSNA Pediatric Bone Age Machine Learning Challenge, as they provide a standardized use case, a common training set, and an objective assessment method applied equally to all models (16). This approach has the benefit of encouraging the development of a diverse set of models and then highlighting not only the best-performing models but also those with the best potential to be combined into high-performing ensembles if the organizers choose to utilize that aspect of the challenge.

Our study had a number of limitations. The challenge test set was used to evaluate the performance of the ensembles and the individual models. By constructing 1000 simulated validation test sets within the original challenge test set to separate model selection from model evaluation, we attempted to limit the effect this might have on our analysis. Furthermore, we were unable to assess generalizability to external datasets, which is of particular importance in medical imaging. Because only model predictions and not the models themselves were submitted, we could not obtain predictions on other data sources. We encourage future competitions to also require model submission to facilitate generalizability studies. In addition, it is likely that many of these submissions incorporated ensemble learning, in which case the ensembles explored in this study actually represent ensembles of ensembles. In fact, all of the top five performing models were actually ensembles themselves (8). In spite of this, we demonstrated that additional performance benefits can be obtained by creating ensembles from unique model ensembles. We speculate that while the ensembles employed by individual competitors reduced primarily computational error, as described previously, these ensembles of ensembles further reduced representational error, given the differences in their underlying approaches. Finally, our analysis only investigated the simple averaging as an ensemble technique. Other techniques for combining predictions, such as stacking, blending, and boosting, were not studied.

In conclusion, we found that by combining machine learning models used to assess the skeletal age on hand radiographs, based on their accuracy and their diversity, ensembles of models could be created that outperformed all individual models. This highlights the importance of both open competition as well as the development of numerous accurate and diverse machine learning models as they are prepared for clinical application. We encourage those conducting similar open competitions in the future to include an analysis of model ensembles in their results.

Acknowledgments: We would like to acknowledge all of the organizers and participants who were involved with the 2017 RSNA Pediatric Bone Age Machine Learning Challenge.

Author contributions: Guarantors of integrity of entire study, I.P., H.H.T., D.B.L.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, I.P., H.H.T., S.S.H., D.L.; clinical studies, I.P., H.H.T., J.K.C, D.B.L.; statistical analysis, I.P., H.H.T., J.K.C., D.L.; and manuscript editing, all authors

Disclosures of Conflicts of Interest: I.P. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: author is consultant for MD.ai. Other relationships: disclosed no relevant relationships. **H.H.T.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: employed by Visiana. Other relationships: disclosed no relevant relationships. **S.S.H.** disclosed no relevant relationships. **J.K.C.** Activities related to the present article: institution receives NIH and NSF grants (NCI: U24 CA180927, NEI: R01EY019474, NSF: SCH1622542). Activities not related to the present article: author is consultant for INFOTECH Soft. Other relationships: disclosed no relevant relationships. **D.B.L.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: advisor for Bunker Hill Health. Other relationships: disclosed no relevant relationships.

References

1. Hansen LK, Salamon P. Neural network ensembles. IEEE Trans Pattern Anal Mach Intell 1990;12(10):993–1001.
2. Ju C, Bibaut A, van der Laan B. The relative performance of ensemble methods with deep convolutional neural networks for image classification. J Appl Stat 2018;45(15):2800–2818.
3. Nanni L, Ghidoni S, Brahnam S. Ensemble of convolutional neural networks for bioimage classification. Appl Comput Inform 2018 Jun 15 [Epub ahead of print] https://doi.org/10.1016/j.aci.2018.06.002.
4. Kahn CE Jr, Kalpathy-Cramer J, Lam CA, Eldredge CE. Accurate determination of imaging modality using an ensemble of text- and image-based classifiers. J Digit Imaging 2012;25(1):37–42.
5. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 2017;284(2):574–582.
6. Norman B, Pedoia V, Noworolski A, Link TM, Majumdar S. Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. J Digit Imaging 2019;32(3):471–477.
7. Dietterich TG. Ensemble methods in machine learning. In: Kittler J, Roli F, eds. International Workshop on Multiple Classifier Systems. Vol. 1857. Berlin, Germany: Springer, 2000; 1–15.
8. Halabi SS, Prevedello LM, Kalpathy-Cramer J, et al. The RSNA Pediatric Bone Age Machine Learning Challenge. Radiology 2019;290(2):498–503.

9. Greulich WW, Pyle SI. Radiographic atlas of skeletal development of the hand and wrist. 2nd ed. Stanford, Calif: Stanford University Press, 1971.
10. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology 2018;287(1):313–322.
11. Gelbrich B, Frerking C, Weiss S, et al. Combining wrist age and third molars in forensic age estimation: how to calculate the joint age estimate and its error rate in age diagnostics. Ann Hum Biol 2015;42(4):389–396.
12. Dunnmon JA, Yi D, Langlotz CP, Ré C, Rubin DL, Lungren MP. Assessment of convolutional neural networks for automated classification of chest radiographs. Radiology 2019;290(2):537–544.
13. Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. PLoS Med 2018;15(11):e1002699.
14. Rodríguez-Ruiz A, Krupinski E, Mordang JJ, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology 2019;290(2):305–314.
15. Tajmir SH, Lee H, Shailam R, et al. Artificial intelligence-assisted interpretation of bone age radiographs improves accuracy and decreases variability. Skeletal Radiol 2019;48(2):275–283.
16. Prevedello LM, Halabi SS, Shih G, et al. Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions. Radiol Artif Intell 2019;1(1):e180031.