RSNA Press Release

Ensembling Improves Machine Learning Model Performance

Released: November 20, 2019

At A Glance

- Combining multiple machine learning models outperforms single-model prediction of pediatric bone age.
- Model ensembles decreased the generalization error of bone age prediction from a mean absolute deviation of 4.55 months to 3.79 months.
- Combining less-correlated and relatively high-performing models resulted in better ensembles than combining the top-performing individual models.

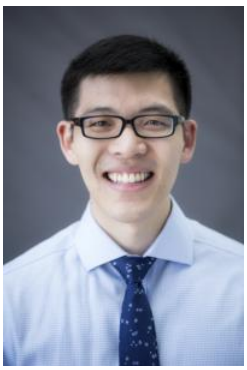
OAK BROOK, Ill. — Ensembles created using models submitted to the RSNA Pediatric Bone Age Machine Learning Challenge convincingly outperformed single-model prediction of bone age, according to a study published in the journal *Radiology: Artificial Intelligence.*

Ensemble learning is a method in machine learning in which different models designed to accomplish the same task are combined into a single model.

Model heterogeneity is an important aspect of ensemble learning. Ensembles tend to perform best when each of the individual models performs well in their own right, and the correlation among individual model predictions is relatively low.

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.

download full-size photo



Ian Pan

“Competitions provide a unique opportunity to study the effects of combining predictions from heterogenous models,” said study author Ian Pan, a medical student at The Warren Alpert Medical School of Brown University in Providence, R.I.

To investigate improvements in performance for automatic bone age estimation that can be gained through model ensembling, Pan and colleagues used 48 submissions from the 2017 RSNA Pediatric Bone Age Machine Learning Challenge.

Participants were provided with 12,611 pediatric hand X-rays 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 X-rays labeled with the weighted average of 6 ratings. The researchers evaluated 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). To estimate the true generalization MAD, they conducted a bootstrap analysis using the 200 test X-rays.

The estimated generalization MAD of a single model was 4.55 months. The best performing ensemble consisted of four models with a 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.

“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,” Pan said. “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 radiological interpretation can be bolstered with multiple readers.

Pan added that the findings also highlight the importance of open competitions like 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.

“Machine learning competitions within radiology should be encouraged to spur development of heterogeneous models whose predictions can be combined to achieve optimal performance,” he said.

For the 2019 RSNA Intracranial Hemorrhage Detection and Classification Challenge, researchers worked to develop algorithms that can identify and classify subtypes of hemorrhages on head CT scans. The data set, which comprises more than 25,000 head CT scans contributed by several research institutions, is the first multiplanar dataset used in an RSNA artificial intelligence challenge.

“Improving Automated Pediatric Bone Age Estimation Using Ensembles of Models from the 2017 RSNA Machine Learning Challenge.” Collaborating with Pan were Hans Henrik Thodberg, Ph.D., Safwan S. Halabi, M.D., Jayashree Kalpathy-Cramer, Ph.D., and David B. Larson, M.D., M.B.A.

Radiology: Artificial Intelligence is edited by Charles E. Kahn Jr., M.D., University of Pennsylvania (Penn) Perelman School of Medicine, Philadelphia, and

Images (JPG, TIF):

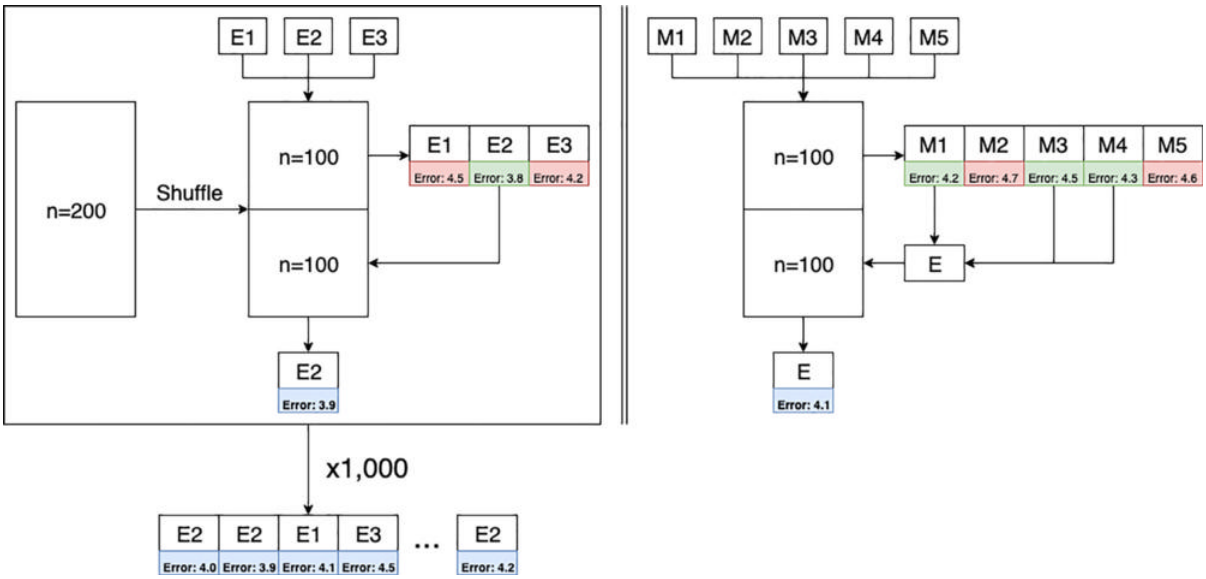


Figure 1. Schematic illustrates the experimental design. E and M refer to theoretical ensembles and individual models, respectively. Left: Ensemble selection and evaluation process of all-model-combinations method where there exist three possible ensembles. The original dataset is randomly split into 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.

High-res (TIF) version

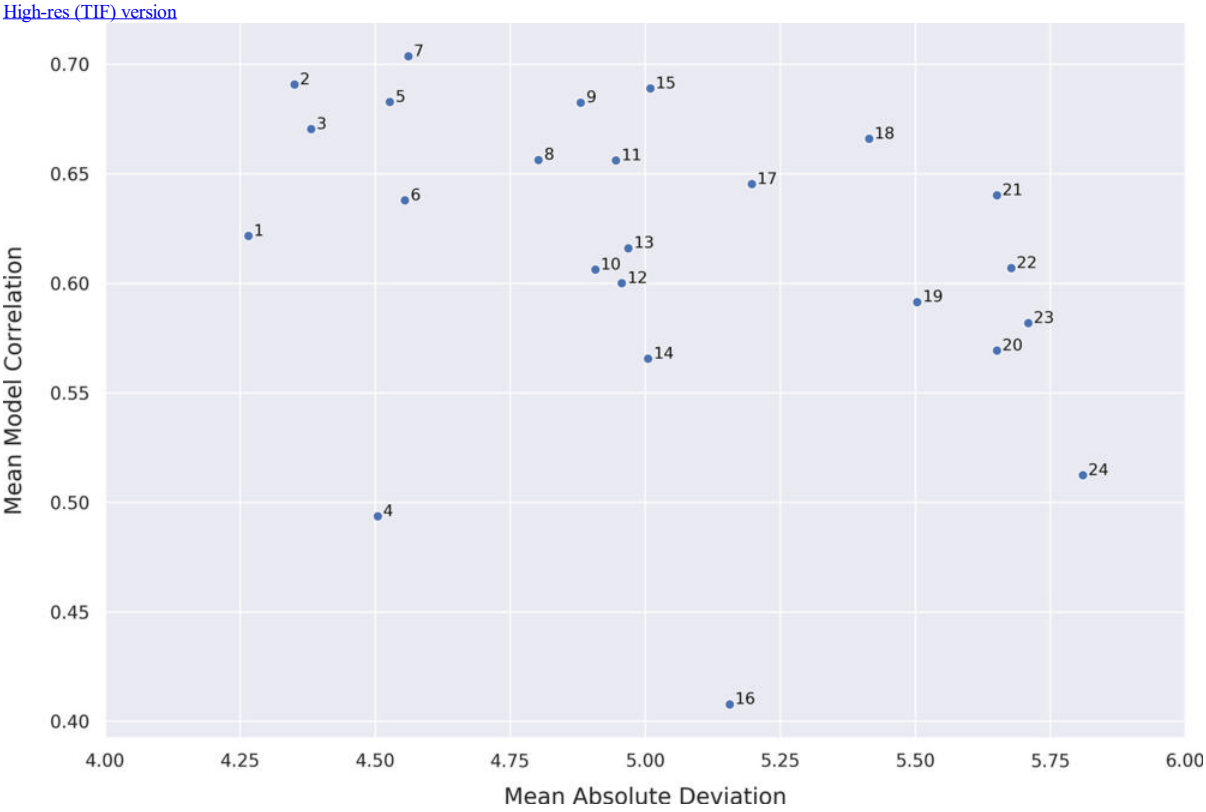


Figure 2. Schematic illustrates the experimental design. E and M refer to theoretical ensembles and individual models, respectively. Left: Ensemble selection and evaluation process of all-model-combinations method where there exist three possible ensembles. The original dataset is randomly split into 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.

High-res (TIF) version

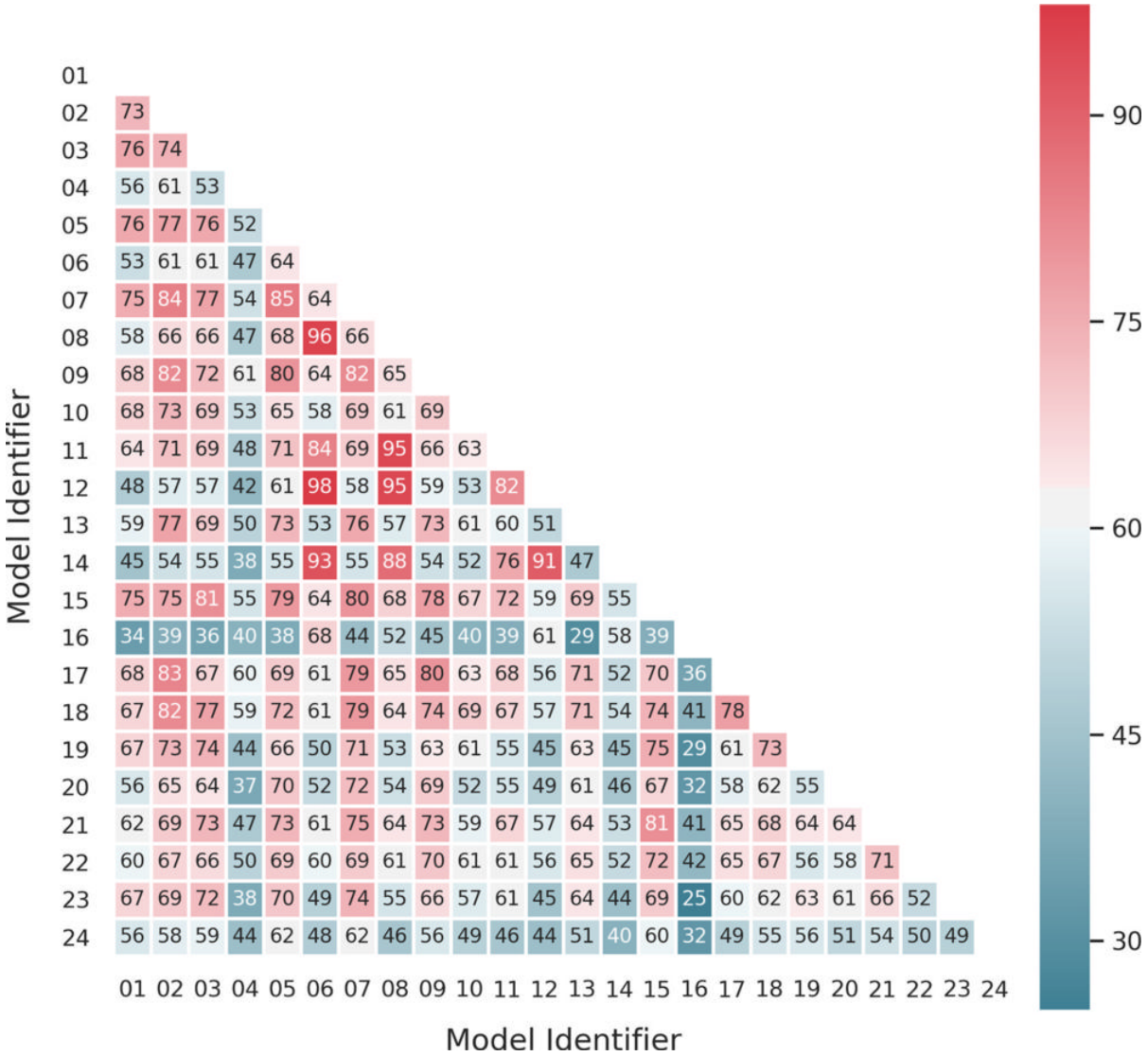


Figure 3. Paired model correlations for pairs of models with mean absolute deviation less than 6 months. Models 4 and 16 stand out as models with low correlation (blue) with other models. Red indicates higher model correlations.

High-res (TIF) version

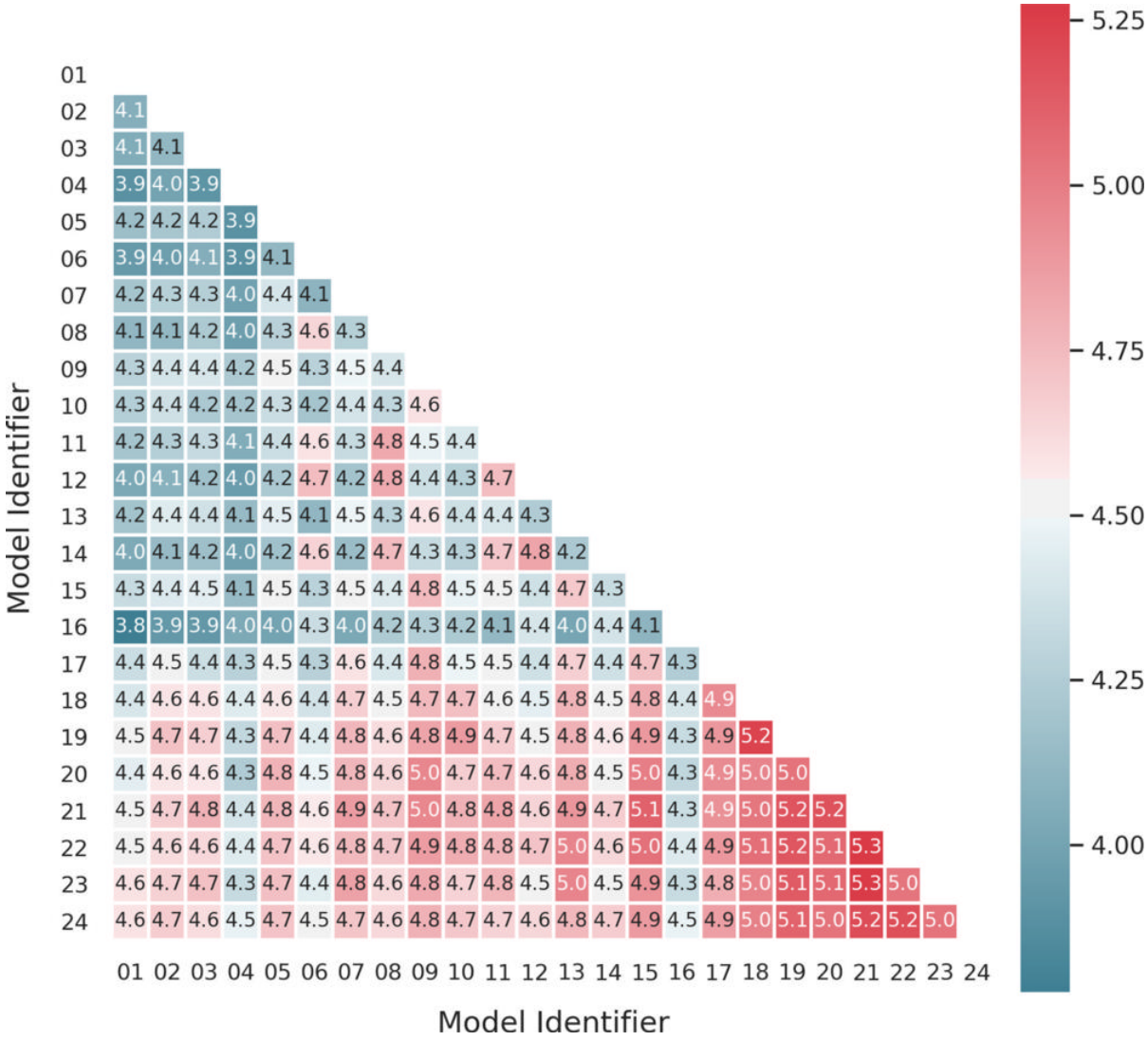


Figure 4. The two-model ensemble mean absolute deviations (MADs) for pairs of models with MAD less than 6 months. MADs are calculated over all 200 cases in the original challenge test set. Red (worse) and blue (better) indicate higher and lower MADs, respectively.

High-res (TIF) version

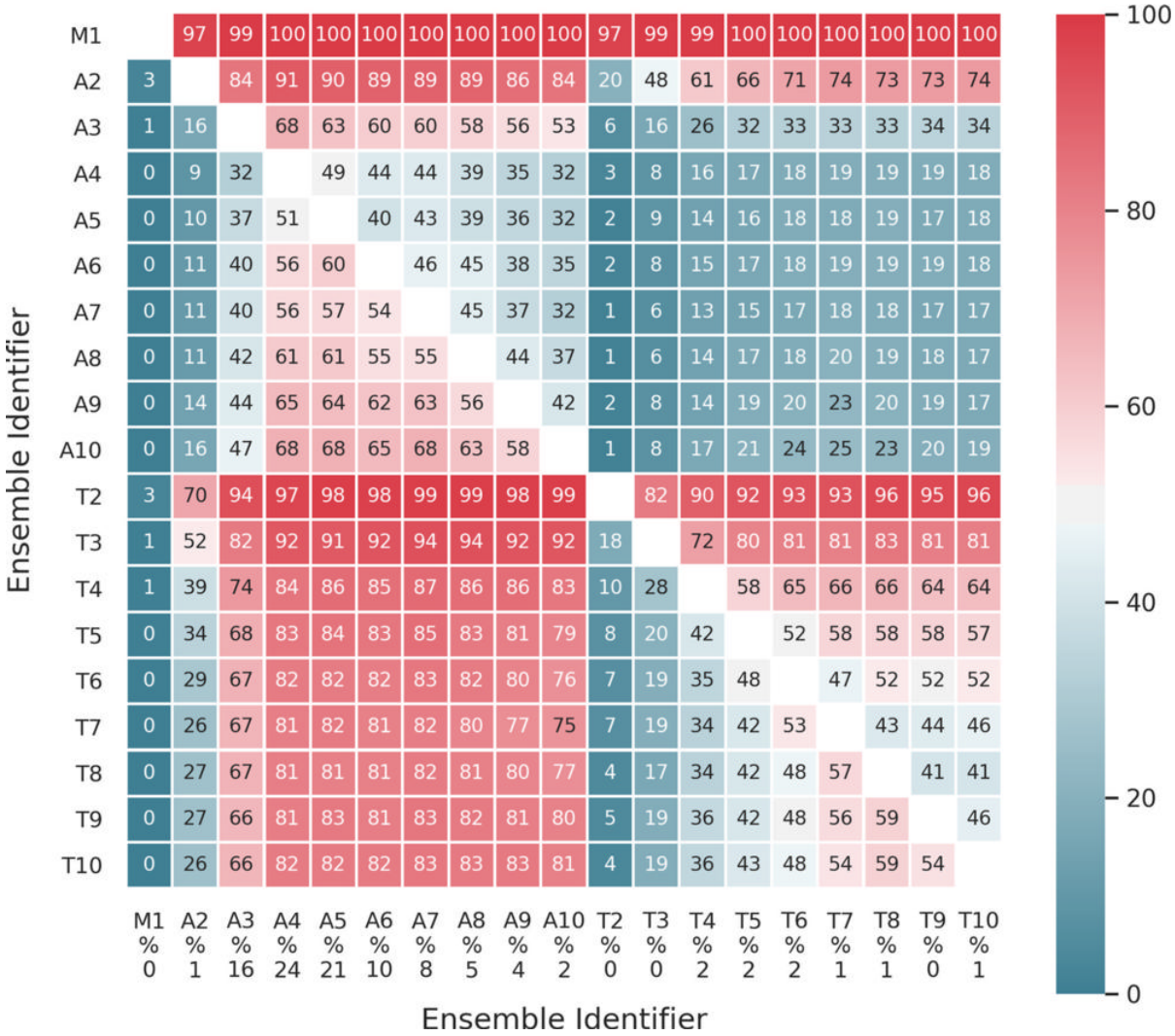


Figure 5. Head-to-head ensemble comparisons across 1,000 experiments. The number displayed is the percentage of experiments where the x-axis ensemble outperformed the y-axis ensemble. M1 represents a single model. A and T refer to all model combinations and top-N ensembles, respectively, and the appended 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.

High-res (TIF) version

Resources:

Editorial
Study abstract