Supplementary materials

Methods:

Greedy ensemble fusion is a supervised machine learning method that can calculate weighted or unweighted averages of multiple predictors to maximize a defined metric of performance. A binary SVM is a supervised machine learning technique that calculates a “hyperplane” (a multi-dimensional boundary) to separate a dataset according to supplied binary labels. The dataset is represented in a multi-dimensional feature space. The output of the SVM is a classification score, which is a signed distance from the hyperplane designating class membership. In comparison to greedy ensemble selection, an SVM is a more complex model, providing opportunity to better fit intricate patterns in data, but at risk to potentially fitting noise in the data.

For score averaging, all algorithm prediction scores on individual images are averaged into a single prediction for that image, with no prior filtering or selection of models. For voting, all algorithm predictions are first dichotomized to values of 0 or 1, using 0.5 as a threshold, and then subsequently averaged for each image. For greedy ensemble fusion, a selection process is employed: algorithm predictions are sorted according to performance, in terms of average precision. An iterative process ensures, whereby for each iteration n, the top n performing algorithm predictions are averaged, and the performance of the average is recorded. The iteration that yields the best overall performance determines which algorithm predictions are selected to be averaged into a single prediction score. For SVMs, feature vectors were created using all participant predictions (sigmoid normalized). A C value of 1 was employed, and thresholds were re-learned according to a probabilistic approach.[22]

Results:

Agreement among dermatologists

The overall kappa for classification and management was 0.53 and 0.47, respectively. Of the 100 lesions, readers were 100% concordant (8/8 agreement) in disease classification of 44 lesions; 26 (59%) were true-positives, 13 (30%) were true-negatives, and 5 (11%) were false-positives. Readers were discordant in diagnosis of 56 lesions, of which 32 were benign and 24 malignant. Regarding the proportion of readers whose disease classification agreed with the reference-standard diagnosis (i.e., histopathology), 8/8 (100%) agreed with the reference-standard diagnosis on 39 lesions, 7/8 (87.5%) on 15 lesions, 6/8 (75%) on 10 lesions, 5/8 (62.5%) on 3 lesions, 4/8 (50%) on 8 lesions, 3/8 (37.5%) on 8 lesions, 2/8 (25%) on 5 lesions, 1/8 (12.5%) on 7 lesions, and 0/8 (0%) on 5 lesions.