Predicting Bone Age of X-Ray Utilizing Deep Neural Networks

Alexander Phu, OMS-I; Alan Lam, OMS-I
Kansas City University College of Osteopathic Medicine

Introduction:

Artificial intelligence (AI) and image recognition provide an opportunity to increase patient care efficiency and reduce physician burnout, particularly in high-stress work settings such as the operating room and emergency department. Although the use of AI and image recognition have been explored in the field of radiology, our program aims to build a machine learning (ML) workflow for learning and identifying bone-age of patients to potentially decrease the imaging workload of the healthcare team (Halabi et. al, 2019).

Methods:

Our team utilized a public dataset sourced from the Radiological Society of North America (RSNA) Pediatric Bone Age Challenge from 2017 to create a ML workflow that could estimate pediatric patient bone ages based on hand radiographs. This consisted of an initial training set containing 12,611 deidentified hand radiographs and a separate test set containing 200 images, including the sex of the patient. An AI algorithm performance was assessed with a correlation coefficient (R^2) as a measure of fit in a regression model using k-fold cross-validation with k=20 to lower bias towards possible data generalization and reduce prediction error. The strongest performing algorithm was then run through a test set to evaluate accuracy in a real-world application. ML workflow design and AI algorithm testing were performed through Orange3 data mining platform.

The training dataset was imported and embedded to extract feature vectors through deep neural networks via Inception v3 embedder. The sample data set we utilized contained two separate files - one file with only PNG images and the other a CSV file containing image names and bone ages. The Merge Data widget was then used to match image feature vectors with image file names and bone ages. We verified the feature vectors and image file names were correctly matched by utilizing the Data Table and Data Sampler widgets to view our training data and ensure accurate importing to prevent output skewing once the testing data set and appropriate regression algorithm were introduced. Using the Test and Score widget, our training data was assessed through kNN, neural network, linear regression, and random forest models. Although the neural network and linear regression yielded a similar R^2 value (Figure 2), it was decided to utilize a linear regression to best generate a single output, predicted bone age, from the various feature vectors. To avoid potential overfitting, the testing data was uploaded as a separate file. Similar to the training data, we utilized the Image Embedding and Merge Data widgets to match the testing image feature vectors to the appropriate test file names and bone ages. The test file was then run with the trained linear regression utilizing the Predictions widget and visualized our results with the Data Table and Scatter Plot widgets.

Results:

Utilizing a linear regression algorithm, our workflow (Figure 1) yielded a R^2 value of 0.799 using training and validation datasets. When tested with a testing dataset, the same linear regression algorithm yielded a R^2 value of 0.698. The training dataset yielded a root mean square error (RMSE) of 18.473 and mean average error (MAE) of 14.300, whereas the testing dataset yielded 23.644 and 18.484, respectively.

Discussion:

The proposed ML workflow based on image recognition has shown potential capability in predicting bone age of pediatric patients using hand radiographs. These programs can be trained to screen for covert adolescent bone abnormalities associated with skeletal maturity (Martin et al., 2011). In future predictive models, a wider diversity in training data should be incorporated for neural network programs to more accurately identify pediatric bone age. A more robust ML workflow can also be considered to account for patient variability in sex, height, and weight.

Figures and Tables:

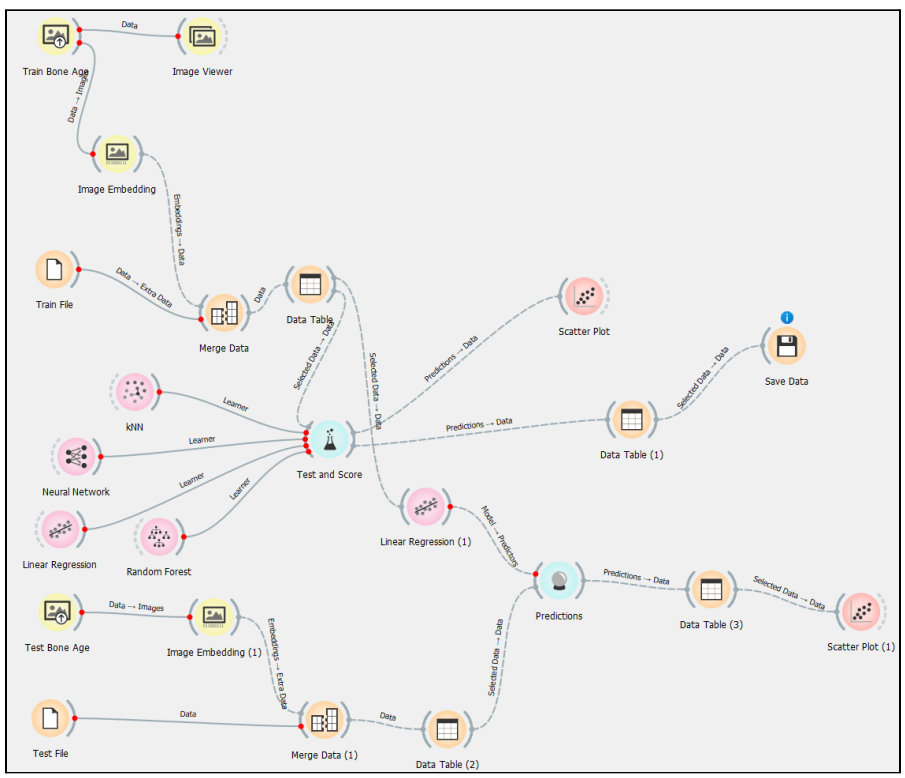
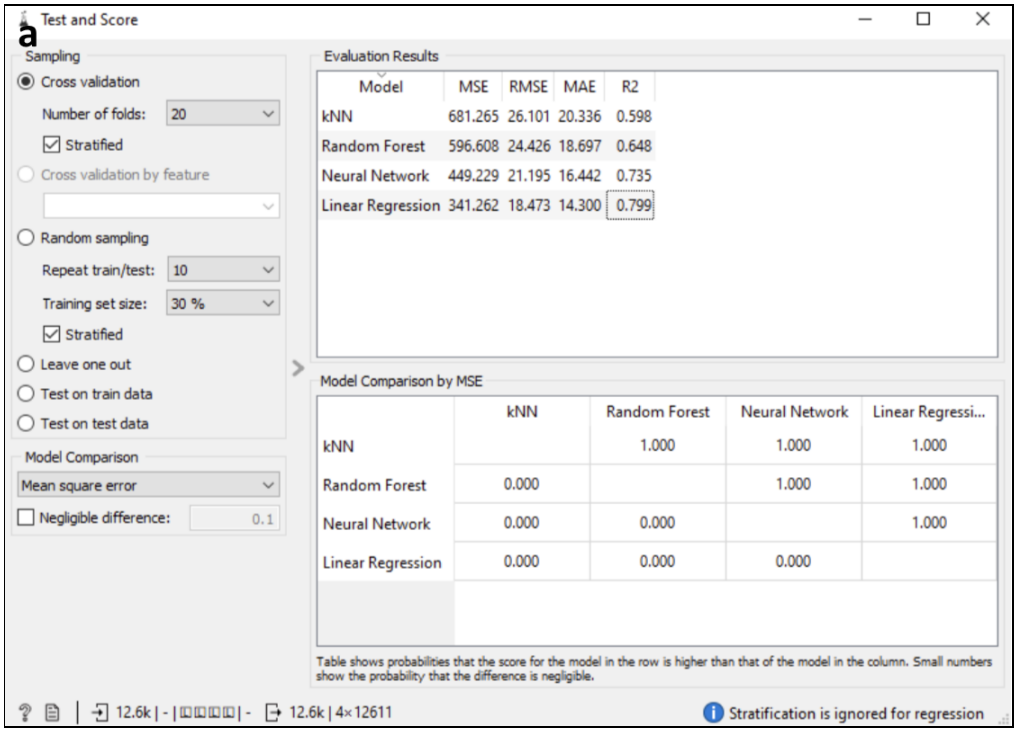


Figure 1. Diagram of the utilized machine learning (ML) workflow on Orange3 data mining platform. Multiple artificial intelligence (AI) algorithms were tested, with linear regression yielding the strongest results with the most accurate predictions when used on testing dataset.



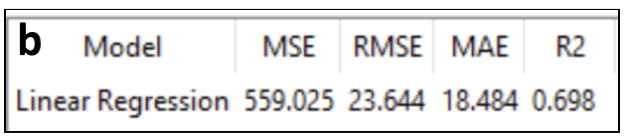


Figure 2. (a-b) Evaluation results from Test and Score widget (a). kNN, Random Forest, Neural Network, and Linear Regression models were tested to evaluate for regression performance. Sampling procedure was set to 20-fold cross validation. Linear regression model then run with testing dataset with Predictions widget (b), yielding evaluation results.

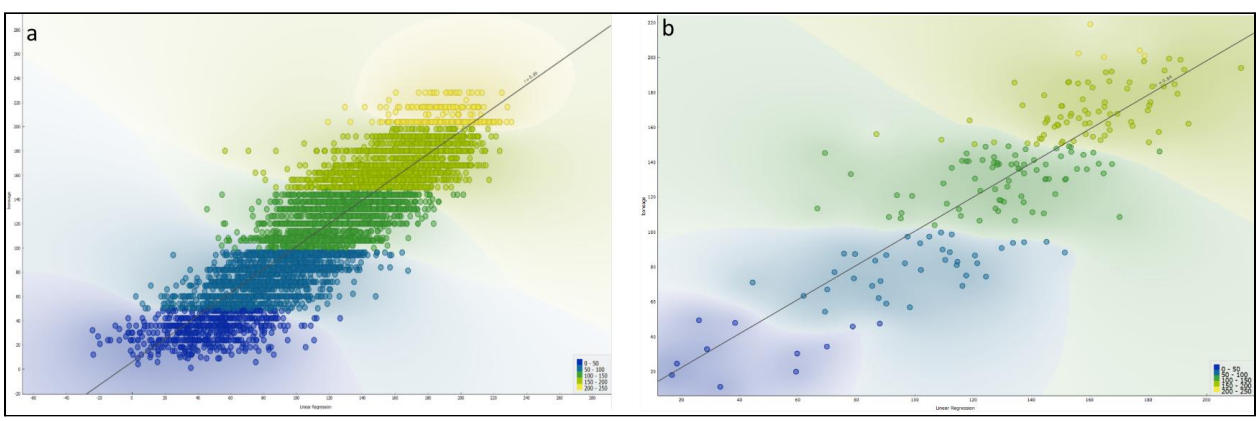


Figure 3. (a-b). Scatter plot of predicted (x-axis) vs. actual bone age (y-axis) of training (a) and testing (b) datasets. Regression lines of scatter plots are r = 0.89 and r = 0.84, respectively.

References:

Halabi, S. S., Prevedello, L. M., Kalpathy-Cramer, J., Mamonov, A. B., Bilbily, A., Cicero, M., Pan, I., Pereira, L.A., Sousa, R. T., Abdala, N., Kitamura, F. C., Thodberg, H. H., Chen, L., Shih, G., Andriole, K., Kohli, M. D., Erickson, B. J., & Flanders, A. E. (2019). The RSNA pediatric bone age machine learning challenge. *Radiology*, *290*(2), 498–503. https://doi.org/10.1148/radiol.2018180736

Martin D, D, Wit J, M, Hochberg Z, Sävendahl L, van Rijn R, R, Fricke O, Cameron N, Caliebe J, Hertel T, Kiepe D, Albertsson-Wikland K, Thodberg H, H, Binder G, Ranke M, B: The Use of Bone Age in Clinical Practice – Part 1. Horm Res Paediatr 2011;76:1-9. doi: 10.1159/000329372