Predicting Appendicular Lean Mass using Supervised and Semi-Supervised Machine Learning Algorithms

Gyaneshwar Agrahari, Kathryn Biles, Alyssa Blount, Tristan Desoto, Berend Grandt, Atif Iqbal, Jacob Kapita

Department of Mathematics, Louisiana State University

Pennington Biomedical Research Center
Louisiana State University

Introduction

Undergraduates and graduate students from the Math department and the Computer Science Department worked jointly with Pennington Biomedical Research Center (PBRC) to find alternatives to dual-energy x-ray absorptiometry (DXA) scans. DXA scans are a costly method of measuring body composition variables such as Appendicular Lean Mass (ALM). These components are important for research in osteoporosis, obesity, and healthy aging (CDC 2021).

Objective

This research aims to predict appendicular lean mass (ALM) cost-effectively. We applied linear regression, polynomial regression, random forest algorithms, neural-network algorithms, support vector regression algorithms, and p-Laplacian methods to predict appendicular lean mass (ALM) efficiently. The primary aim is to conduct a comparative analysis of these methodologies, evaluating their predictive performance using error analysis and comparing the true versus predicted values of the models. To accomplish this, we utilized the root mean squared error (RMSE) to evaluate the alignment of the model's average predictions with the actual average values, based on specific parameter settings. Our results are displayed to the right.

Methodology for Supervised Learning

Pennington provided a dataset encompassing information on 846 patients, featuring age, which ranged from five to eighty-nine years old, sex, race, and forty-four biomarkers along with DXA-measured ALM values.

Data Pre-processing: For those patients for whom we did not have information for all biomarkers, we ignored them in our analysis.

Testing: To evaluate model accuracy in supervised learning, we created a test set by randomly choosing 20 percent of all males and 20 percent of all females(49 males and 54 females). The rest of the males and females were part of the training set.

Supervised Machine Learning Methods:

1.) Linear/Polynomial Regression This method is a statistical approach used to model the relationship between the ALM and the biomarkers by fitting a linear or polynomial equation of higher degree to the observed data.

2.) Random Forest A random forest is a supervised learning algorithm that combines the predictions from multiple decision trees to produce more accurate and stable predictions.

3.) Neural-Networks Our Neural network is a Multilayer Perceptron (MLP) model, comprising two hidden layers. The input layer uses 39 relevant biomarkers from the data. We use a sigmoid activation layer between the first hidden layer and the second hidden layer and between the second hidden layer and the output. Since this is a regression problem, we do not apply the activation function to the output.

4.) Support Vector Regression Support Vector Regression (SVR) is a type of machine learning model that uses the principles of support vector machines to predict continuous outcomes. It works by fitting the best line (in a higher or infinite dimensional space) within a threshold error margin, efficiently handling both linear and non-linear data.

Results for Supervised Learning

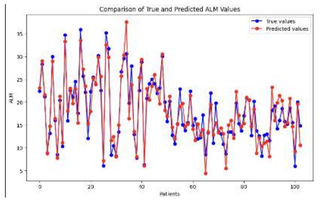

Figure 1. True vs Predicted ALM of Polynomial Regression for degree=2

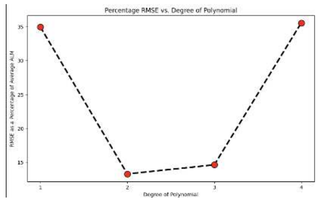

Figure 2. Error Analysis of Polynomial Regression (including Linear)

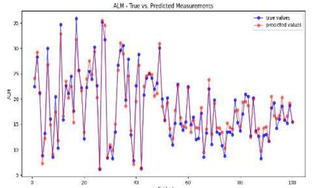

Figure 3. True vs Predicted ALM of Random Forest for Number of Trees=50

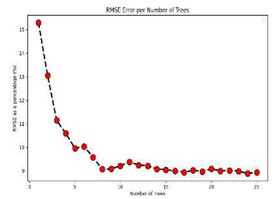

Figure 4. Error Analysis of Random Forest

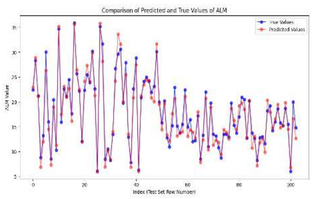

Figure 5. True vs Predicted ALM of Neural-Networks for 350 epochs

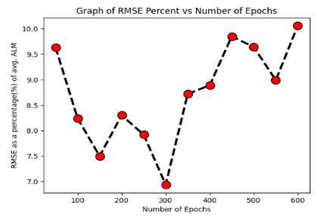

Figure 6. Error Analysis of Neural-Networks

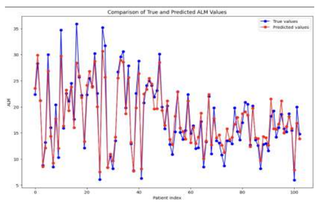

Figure 7. True vs Predicted ALM of Support Vector Regression with Kernel Function RBF and $\epsilon = 0.8$

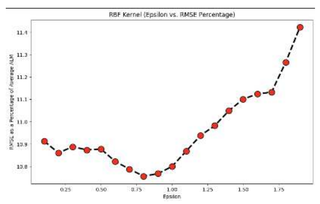

Figure 8. Error Analysis of Support Vector Regression with Kernel Function

Methodology for Semi-Supervised Learning

In a 2-Laplacian-based semi-supervised learning algorithm, we represent the data as a graph. The nodes represent the patients and two nodes are connected if the corresponding patients are similar enough in their characteristics. The next step is to divide the nodes into two categories: Labeled and Unlabeled. We then apply a biased random walk starting at an unlabeled node until we reach a label node. This process is repeated to predict the label of the unlabeled node we started with.

Results for Semi-Supervised Learning

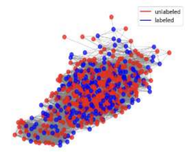

Figure 9. Patient Network for $\epsilon = 0.8$ and Label Rate=40 %

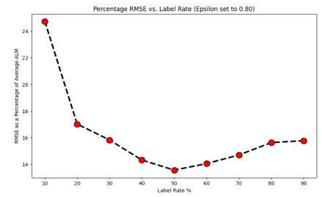

Figure 10. Error Analysis of 2-Laplacian

Future Work

Future studies will build upon these models to predict various DXA measurements, such as bone density and body fat percentage. Additionally, we plan to extend the 2-laplacian algorithm to a generalized p-Laplacian for $p \geq 2$. We hope that our models will be integrated into future obesity, metabolism, and nutrition studies.

Acknowledgements

- We would like to thank Professor Peter Wolenski and Dr. Nadejda Drenska from the Department of Mathematics, Louisiana State University, for putting together the whole team and helping make this project a success.
- Department of Mathematics, Louisiana State University.
- Pennington Biomedical Research Center, LSU.

References

- Center for Disease Control and Prevention. (2021, October 20). *Radiation in Healthcare: Bone Density (DEXA Scan).* https://www.cdc.gov/nceh/radiation/dexa-scan.html