An Initial Evaluation of Different Machine Learning Algorithms in Predicting Hypertension in African Populations Using Biochemical and Physical Activity Features

Brendyn Miller[1], Anamul Haque[2], Dr. Will Richardson[1]

[1]Department of Bioengineering, Clemson University
[2]Department of Biomedical Data Science and Informatics, Medical University of South Carolina
February 10th, 2022 at 3:30 PM

Introduction: Machine learning (ML) models have shown significant promise in accurately predicting the onset of disease and in aiding physicians in making clinical decisions to provide their patients with optimal care.[1] When implemented correctly, ML models can improve diagnoses, aid in risk evaluation, and help in selecting optimal treatments and therapies.[2] However, ML models are only as viable as the data used to train the model and have the drawback of being difficult to interpret from a clinical perspective. Therefore, it is crucial that the data used to train ML models is relevant to the prediction being made with the model and that algorithms are implemented which can enable physicians to understand the importance of the factors which led to the model's prediction. The goal of this study was to build a ML model pipeline that could evaluate both biochemical and biomechanical variables to determine whether a patient would be considered at high risk for developing hypertension and the incorporating algorithms into that model which would allow users to assess which factors were most relevant to the development of hypertension.

Methods: To develop our initial ML model pipeline, we used data collected from the African-PREDICT study, a research initiative which collected a large range of data with 420 different features from 1,262 patients located in the northwest province of South Africa.[3] From this dataset, we selected blood serum biomarkers and physical activity measures (38 variables total) as input features for our model. K-nearest neighbors (KNN) imputation was used to impute missing variables in the dataset and MinMax Scaling was applied to standardize the data. Feature selection techniques were applied to filter out variables that were redundant or irrelevant to the model. A variety of different ML algorithms (Random Forest, Support Vector Machine, KNN, XGBoost) were then trained on the prepared dataset and accessed by evaluating the model's accuracy and F1 score. Cross-validation was used to reduce model overfitting and selection bias towards the training data and grid hyperparameter search algorithms were implemented to optimize model performance. Finally, feature importance was assessed using a random forest classifier algorithm to determine which features were most relevant to the model prediction.

Results: An initial evaluate of a trained random forest ML model demonstrated an accuracy of 91% and F1 score of 0.95%; however, model evaluation is currently ongoing. In conclusion, an initial ML model pipeline has been successfully developed, allowing for different ML algorithms to be implemented easily. Future research is being conducted to determine methods to optimize the model and improve metric accuracy. Immediate next steps include research into the optimal feature selection technique and other options for assessing feature importance.

References:

1) Kohli, P. S., & Arora, S.IEEE. 2018 | 2) D. Maddux, Acumen Physician Sol. 2014 | 3) Schutte AE, Gona PN, Delles C, et al. J Prev Cardiol. 2019