Title: Machine learning-based risk factor analysis and prevalence prediction of intestinal parasitic infections.

Presenter: Ahmet Ay, Associate Professor of Biology and Mathematics, Colgate University, Hamilton, NY

Authors: *Zafar A, Attia Z, Tesfaye M, Walelign S, Wordofa M, Abera D, Desta K, Tsegaye A,* ***Ay A,*** *Taye B.*

Abstract:

Background: Previous epidemiological studies have examined the prevalence and risk factors for a variety of parasitic illnesses, including protozoan and soil-transmitted helminth (STH, e.g., hookworms and roundworms) infections. Despite advancements in machine learning for data analysis, the majority of these studies use traditional logistic regression to identify significant risk factors.

Methods: In this study, we used data from a survey of 54 risk factors for intestinal parasitosis in 954 Ethiopian school children. We investigated whether machine learning approaches can supplement traditional logistic regression in identifying intestinal parasite infection risk factors. We used feature selection methods such as InfoGain (IG), ReliefF (ReF), Joint Mutual Information (JMI), and Minimum Redundancy Maximum Relevance (MRMR). Additionally, we predicted children's parasitic infection status using classifiers such as Logistic Regression (LR), Support Vector Machines (SVM), Random Forests (RF) and XGBoost (XGB), and compared their accuracy and area under the receiver operating characteristic curve (AUROC) scores. For optimal model training, we performed tenfold cross-validation and tuned the classifier hyperparameters. We balanced our dataset using the Synthetic Minority Oversampling (SMOTE) method. Additionally, we used association rule learning to establish a link between risk factors and parasitic infections.

Key Findings: Our study demonstrated that machine learning could be used in conjunction with logistic regression. Using machine learning, we developed models that accurately predicted four parasitic infections: any parasitic infection at 84.7% accuracy, helminth infection at 89.3%, STH infection at 95.9%, and protozoan infection at 95.0%. The XGBoost (XGB) classifier achieved the highest accuracy when all risk factors were considered. The best predictors of infection were socioeconomic, demographic, and hematological characteristics. Despite its mitigating effect on STH infections, mass deworming was also associated with increased protozoan infections.

Conclusions: We demonstrated that feature selection and association rule learning are useful strategies for detecting risk factors for parasite infection. Additionally, we showed that advanced classifiers might be utilized to predict children's parasitic infection status. When combined with standard logistic regression models, machine learning techniques can identify novel risk factors and predict infection risk.