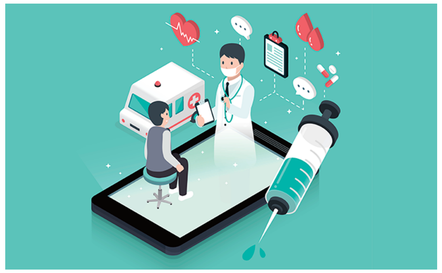
Leverage Electronic Health Record to Advance Artificial Intelligence Risk Prediction

Yirui Hu[1], Kunpeng Liu[2], Nana Wang[3], H Lester Kirchner[1] (yhu1@geisinger.edu)

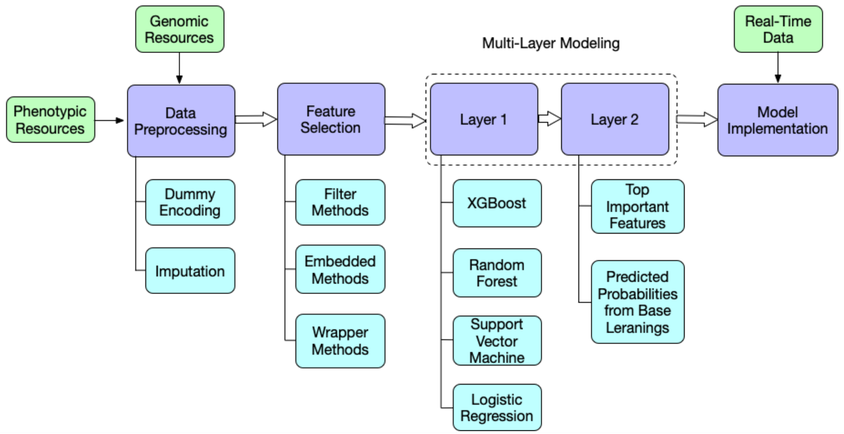
[1]Biomedical and Translational Informatics, Geisinger; [2]Computer Science, Missouri S&T; [3]Biometrics Research, Merck

BACKGROUND



- Increased clinical use of electronic health record (EHR) is providing massive amounts of patient data.
- The development of predictive analytics using machine learning techniques is essential to provide insights and prediction for clinically-relevant outcomes using historical EHR data.
- We present an **EHR predictive analytics pipeline** to process massive amounts of historical EHR data and then implement it for risk assessment for new patients.

Framework



METHODS

Key steps in the predictive analytics pipeline:

Step 1: Data preprocessing. Input features from phenotypic and genomic resources are normalized for continuous features and categorized for categorical features with dummy coding. Missing data will be imputed using Multivariate Imputation by Chained Equations.

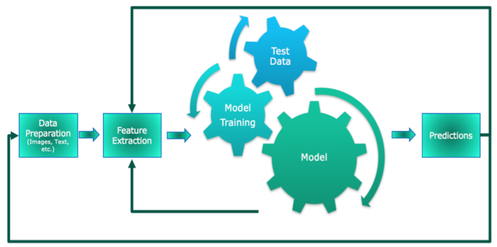
Step 2: Feature selection. Feature selection techniques including filter methods, wrapper methods and embedded methods, are adopted to determine the optimal feature subset. The selected feature subset will substitute the original feature set in the subsequent modeling steps.

Step 3: Multi-layer modeling. Layer 1: Train base learners with cross-validation. The base learner includes logistic regression, support vector machine, random forest, and eXtreme Gradient Boosting (XGBoost). The parameters for each algorithm are tuned by cross-validated search over a fixed number of parameter settings.

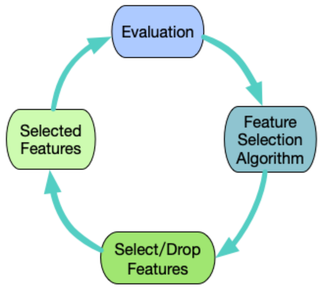
Layer 2: Combine the top important features with the predicted probabilities from base learners, and put them into XGBoost to get the final prediction.

Step 4: Model implementation. To further implement the artificial intelligence in real-time data. The top importance features will be monitored in live data input.

A Standard Machine Learning Pipeline



A Standard Feature Selection Pipeline



CONCLUSIONS

The predictive analytics pipeline we developed can risk stratify patients to different tiers of risk, facilitate early identification of high-risk patients and help optimize care delivery in large health care system. Future work will be focused on validating the prediction algorithms externally, or across in multiple sites for maximally portable and generalizable artificial intelligence algorithms that can be widely adopted in diverse clinical settings.

REFERENCES