An explainable XGBoost–based approach towards Preeclampsia risk prediction

Delwende E Birba

Abstract

Preeclampsia (PE) is a heterogeneous and complex disease associated with rising morbidity and mortality in pregnant women. Early recognition of patients at risk is a pressing clinical need to significantly reduce the risk of adverse pregnancy outcomes. This study aimed to develop and evaluate an explainable machine learning for preeclampsia prediction. Extreme gradient boosting method were used to construct the prediction models and achieved AUC score of 0.90.

Introduction

Preeclampsia (PE) is a pregnancy-specific syndrome that affects 3-5% of pregnant women and is characterized by hypertension, and proteinuria [1]. Many studied have applied ML techniques that include metabolites, images analyses, and risk factors datasets, among others to diagnose and to predict PE. However, they do not present any elements to support Explanations ability for medical experts. **Aim:** To develop models using machine learning to predict late-onset preeclampsia based on plasma protein.

Methodology

eXtreme Gradient Boosting (XGBoost) were used with 5 folds cross validation to train the data. eXtreme XGBoost constitutes an efficient and scalable variant of the Gradient Boosting Machine (GBM) algorithm, leveraging the power of decision tree ensembles towards performance optimization.
Interpretability: SHAP values for XGBoost interpretability

Results

DATA

- Public dataset used in [3]
- 90 patients with normal pregnancies (controls) and
- 76 patients with late-onset preeclampsia defined as preeclampsia diagnosed at ≥34 weeks of gestation)
- **1125 proteins**

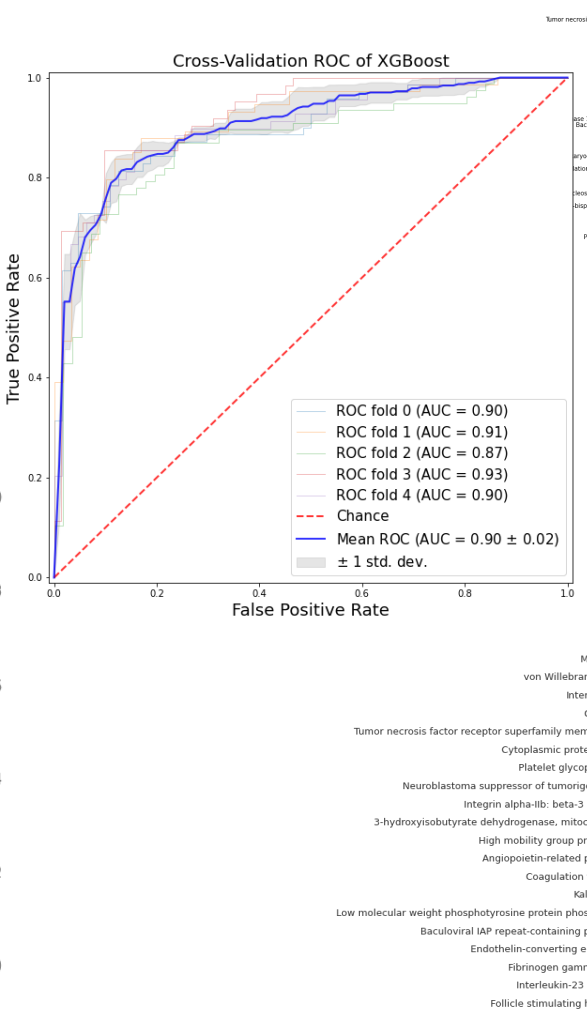
Total features: 1230

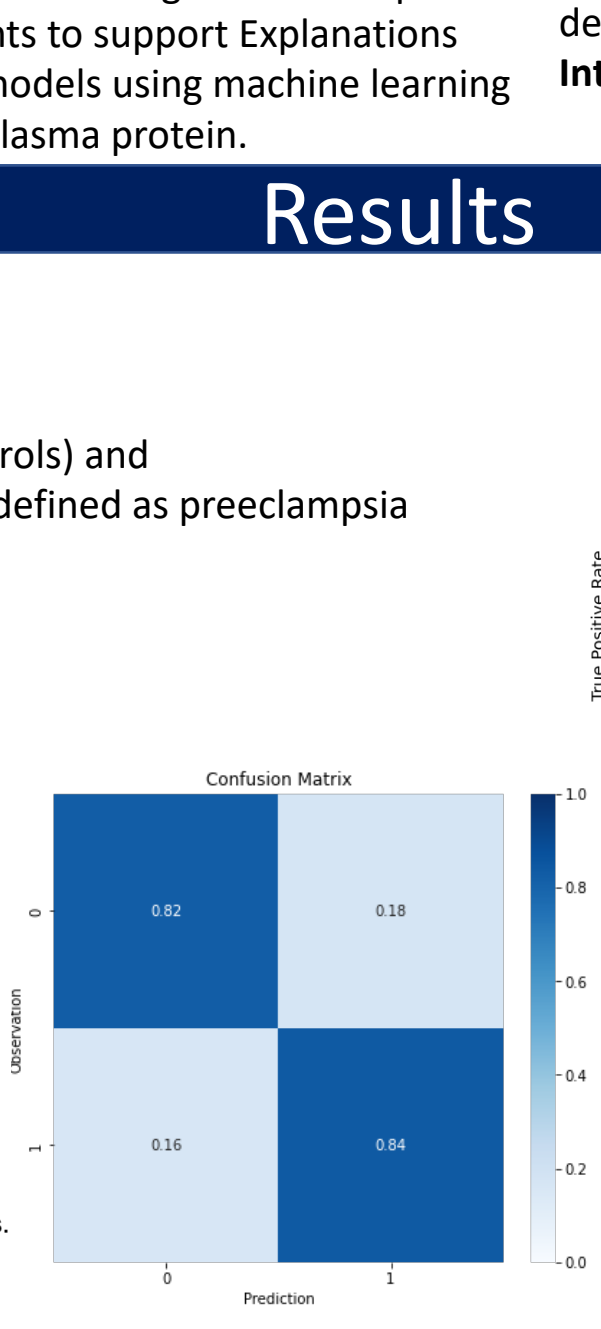
Sample size: 666 samples

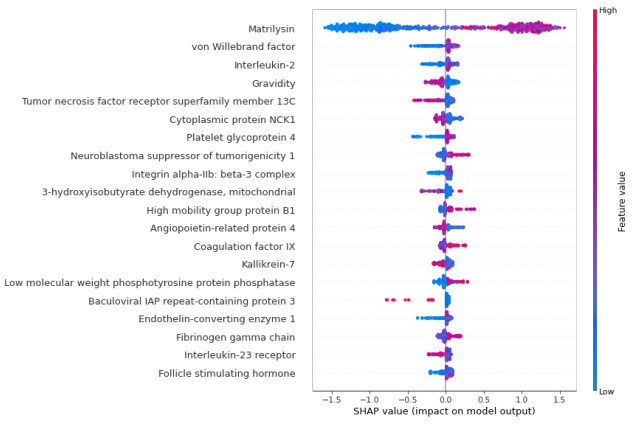
Table: Patient and pregnancy data.

Characteristics	Preeclampsia (n=76)	Controls (n= 90)
Age (Years)	22.0 [21.0–29.0]	24.0 [21.0–27.8]
Body max index (kg/m2)	30.0 [24.8.0–36.2]	26.5 [22.8–33.2]
Race		
African American	72	84
Caucasian	4	6
Pregnancy		
Parity((nulliparity)	32	26
Gravidity	2[1-4]	3[2-5]
Gestational age at delivery	38.7 [37.7–39.4]	39.4 [39.0–40.4]

Reported are the median and interquartile range or absolute numbers.







Conclusion

Our results suggest that explainable machine-learning models can be used to identify women at risk for preeclampsia, which in turn, can improve the development of preventive strategies. Our models achieved AUC score of 0.90.

Reference

[1] A. Filipek and E. Jurewicz, “Preeclampsia - a disease of pregnant women,” Postepy Biochemii, vol. 64, no. 4, pp. 229–232, 2018.

[2] T. Chen and C. Guestrin, “XGBoost,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016.

[3]https://figshare.com/articles/dataset/Proteomic_Models_in_Preeclampsia/7962998/1

HOST INSTITUTIONS

University College Dublin
An Coláiste Ollscoile, Baile Átha Cliath

FUNDED BY: