A Cluster-based Synthetic Cohort Framework Supports Generalisable 4 Prediction of Celiac Disease Severity from Routine Data

Vara Dutt
British School of Jakarta, Jakarta, Indonesia

Michael Bryan
Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom

Abstract:

Purpose: Celiac disease (CD) remains underdiagnosed in low-resource settings, where access to diagnostic serology such as tTG-IgA is limited or unavailable. This presents a particular challenge for high-risk groups like individuals with diabetes, in whom delayed diagnosis increases the risk of complications. We developed CD-SENSE, a non-invasive machine learning tool that predicts histological severity (Marsh classification) using routine clinical features. Our aim was to create a scalable screening solution that does not depend on CD-specific serological tests.

Methods: We used a publicly available dataset from Wageningen University consisting of IgA, IgG, symptom indicators, and demographic features. Four analytical pipelines were constructed to test the impact of feature categorisation and class imbalance. Multiple classifiers were tuned via grid search and evaluated using 1,000 bootstrap iterations across accuracy, AUC, precision, recall, and F1-score. A held-out validation set (20%) was isolated prior to feature selection and training to simulate external generalisation. We further developed a cluster based synthetic stress test to evaluate robustness under biologically plausible, unseen patient scenarios. Using k-means clustering (k=10) on z-scored IgA/IgG and binary symptoms, we generated synthetic samples via multivariate normal sampling and Bernoulli draws, constrained by cluster-specific distributions and biological plausibility. Clusters with majority-class purity ≥0.55 were retained.

Results: XGBoost consistently outperformed other models, achieving 92.6% accuracy and an AUC of 0.974 on the held-out validation set. The most predictive features were total IgA, IgG, short stature, and weight loss. In the synthetic stress test, CD-SENSE retained an accuracy of 81.7%, demonstrating resilience to novel but clinically realistic input combinations. Model interpretability via permutation importance and SHAP confirmed biologically consistent feature rankings. CD-SENSE has been deployed as a freely accessible web application.

Conclusion: CD-SENSE enables cost-effective, serology-free screening for celiac disease severity using data routinely available in clinical encounters. This may be particularly impactful in resource-