SIIM CAIMI25
Conference on Artificial Intelligence in Medical Imaging
OCT 20-21 | UNIVERSITY OF CALIFORNIA, IRVINE

Foundation Model-Based Semi-Supervised Multiple Instance Learning for Automated Barrett's Esophagus Dysplasia Classification

Amirali Khosravi, MD, Mayo Clinic; Jaidip M. Jagtap, PhD; Bradley J. Erickson, MD, PhD, CIIP, FSIIM; Prasad G. Iyer, MD, FACG; Chamil Codipilly, MD

Introduction/Background

Barrett's esophagus (BE) dysplasia diagnosis suffers from substantial interobserver variability due to subjective criteria and inflammation-induced artifacts. This leads to overcalls of dysplasia grade particularly in the community. We aimed to develop a deep learning model using a large multi-center cohort to improve detection of BE dysplasia grade.

Methods/Intervention

We utilized 969 whole slide images (WSIs) from four academic centers (Mayo Clinic, Baylor, OHSU, UAB) collected from January 1992 to September 2022. All patients had endoscopic BE (≥1 cm columnar mucosa) with confirmed intestinal metaplasia. Ground truth was established by consensus review of two expert GI pathologists. We employed a semi-supervised training approach using annotated regions from 465 slides where pathologists marked 1,990 regions of interest containing the highest dysplasia grade. WSIs were divided into 1280×1280 overlapping patches with features extracted using the Midnight pathology foundation model. We first trained three patch-level binary classifiers on the annotated regions, then used these to select the top 400 candidate patches per WSI across all centers, sampling high-confidence predictions from each dysplasia grade in fixed proportions. An attention-based multiple instance learning model (AttriMIL) was trained to predict WSI-level dysplasia grade. Data were stratified by dysplasia grade and center, split 70-15-15 for training, validation, and testing. Model performance was evaluated using AUC and F1 score.

Results/Outcome

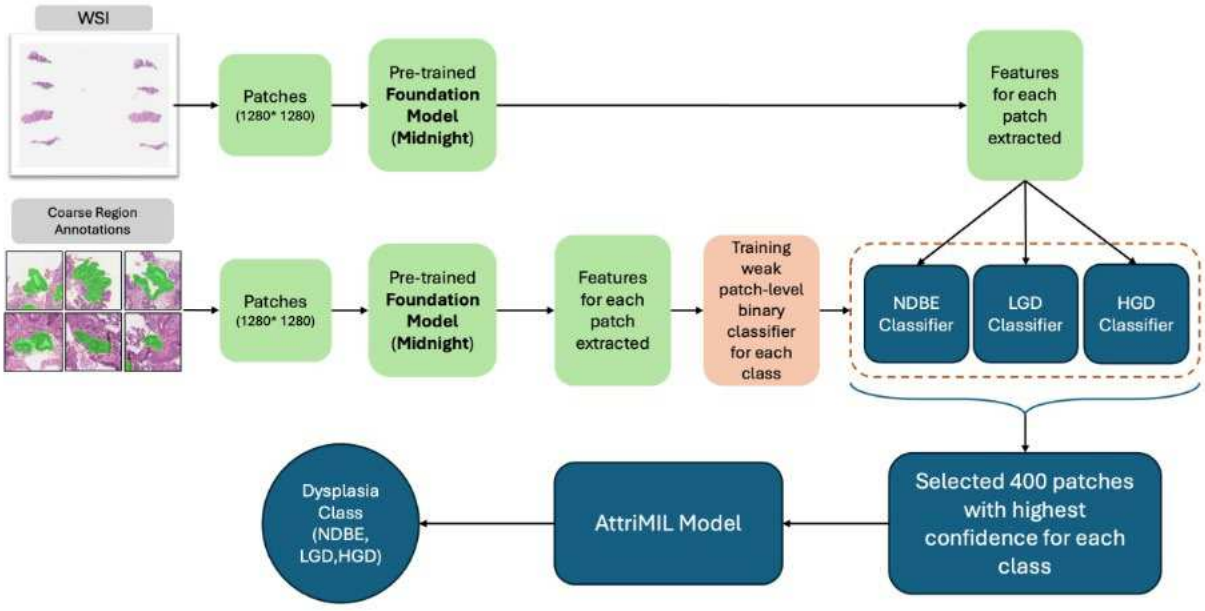
The full cohort had mean age of 64.6 ± 12.8 years with 81.9% being men. Dysplasia grade included 365 nondysplastic BE (NDBE), 319 LGD, and 285 high-grade dysplasia (HGD). On the test set, the model achieved AUC of 95%, F1 score of 82%, sensitivity of 82%, and specificity of 91% for three-class grading. For binary classification of any dysplasia versus NDBE, AUC was 97%, F1 score 93%, sensitivity 90%, and specificity 94%.

Conclusion

Our semi-supervised approach achieved excellent performance for BE dysplasia grading. The combination of foundation models and intelligent patch sampling effectively leveraged limited annotations across a large multi-center dataset, providing robustness without center-specific adaptations.

Statement of Impact

This model may serve as a valuable diagnostic adjunct, reducing interobserver variability and improving consistency in Barrett's esophagus dysplasia assessment, particularly benefiting community pathologists where overcalls are common.



Summary of pre-processing and training of Barrett's Esophagus Dysplasia Classifier

	Sensitivity (%)	Specificity (%)	F1 score (%)	AUC (%)
3-class results				
NDBE	94	89	88	95
LGD	63	94	72	
HGD	89	89	82	
2-class results				
NDBE	94	89	88	97
Dysplastic BE	90	94	93	

Model Performance Metrics for Barrett's Esophagus Dysplasia Classification

Keywords

Barrett's esophagus; Semi-supervised learning; Multiple instance learning; Foundation models; Digital pathology