Detecting Manuscripts Related to Computable Phenotypes Using a Transformer-based Language Model

Junghoon Chae, PhD[1], David Heise, MS[1], Keith Connatser[1], Jacqueline Honerlaw, RN, MPH[2], Monika Maripuri, MBBS, MPH[2], Yuk-Lam Ho, MPH[2], and Kelly Cho, MPH, PhD[2,3]

[1]Oak Ridge National Laboratory; [2]VA Boston Healthcare System; [3]Mass General Brigham, Harvard Medical School

U.S. Department of Veterans Affairs
VA Boston Healthcare System
VA Atlanta Healthcare System

Background & Significance

- Building a phenomics library requires identifying manuscripts that define computable phenotypes.
- Manual screening of biomedical literature is time-consuming, inconsistent, and unscalable.
- Transformer-based NLP models can automate classification, but their 512-token limit restricts full-text processing.
- **A scalable, domain-specific NLP model capable of classifying long biomedical manuscripts is needed.**

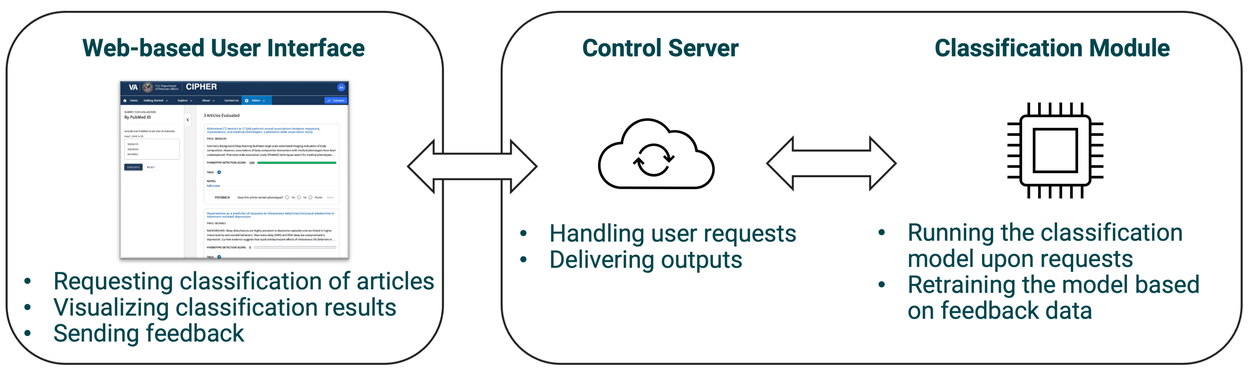
Objective

- Develop a transformer-based classification model that identifies phenotype-relevant manuscripts.
- Integrate the model into the Centralized Interactive Phenomics Resource (CIPHER)[1] platform for automated, feedback-driven curation.

Methods

Systemic Framework for Computable Phenotype Literature Curation.

- Web-based User Interface for submitting articles and viewing results.
- Control Server for managing requests and responses.
- Classification Module for running the model and retraining based on user feedback.



Classification Model Development

- Developed through several iterative stages.
- Began with traditional machine learning algorithms suited to the initial data size and characteristics.
- Transitioned to more advanced models.

Transformer-based Sliding-Window Model

- Fine-tuned BioBERT on annotated manuscripts.
- Implemented a sliding-window segmentation to handle > 512-token count documents.
- Applied token-count–weighted aggregation to generate final document-level predictions.

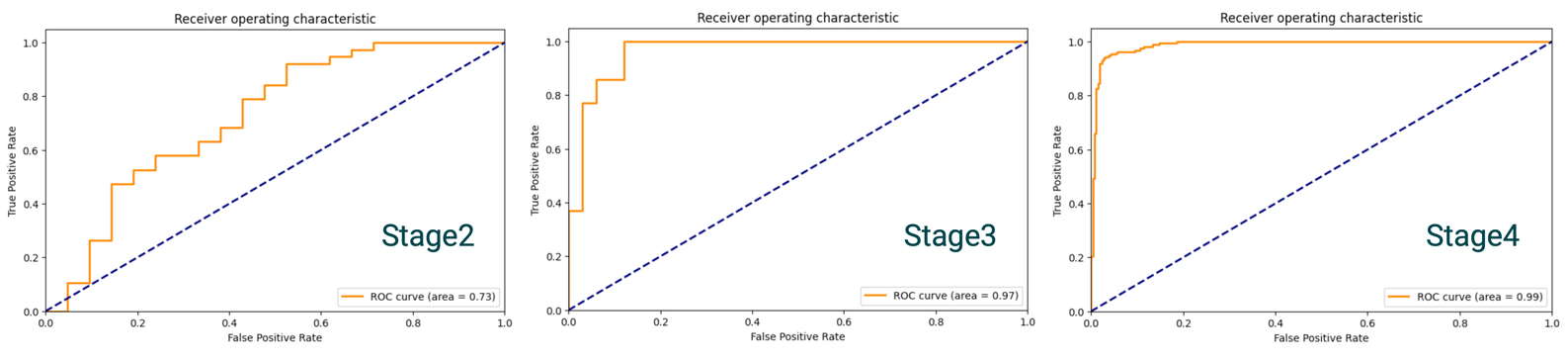
Results

- The classification module of the framework was developed through several iterative stages.
- As the dataset expanded and the documents required deeper contextual and semantic understanding, we transitioned to more advanced models.
- Each stage achieved notable improvements in classification performance, as summarized in the table below.

Stage	Technique Name	Data (number of manuscripts)	Accuracy (%)
Stage 1	Random Forest	145	60
Stage 2	BioBERT	145	72
Stage 3	BioBERT	226	88
Stage 4	**BioBERT with Sliding-Window**	**392 (3,571 512-token segments)**	**95**

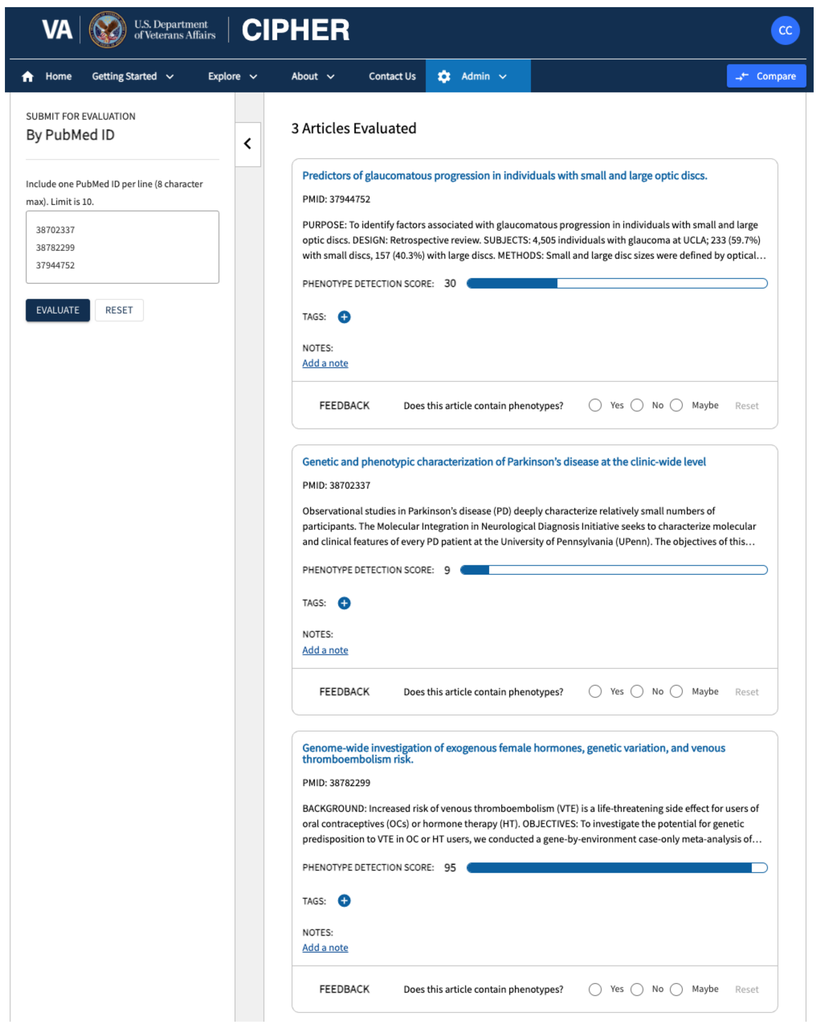
Receiver Operating Characteristic (ROC) Curves for Model Development Stages

- Stage 2: Transformer-based model with original data, Area Under Curve (AUC) = 0.72.
- Stage 3: Transformer--based model with balanced data, AUC = 0.88.
- Stage 4: BioBERT with Sliding-Window approach, AUC = 0.99.
- The ROC curve for this iteration clearly shows a further upward and leftward shift, indicating enhanced sensitivity and specificity, and a more robust classification capability due to the improved data quality.



Use Case: CIPHER (https://phenomics.va.ornl.gov/)

- Submit PubMed IDs for article classification.
- View predicted phenotype detection scores.
- Provide feedback (Yes/No/Maybe) to continuously refine the underlying model.



[1] Honerlaw, J., et al., Centralized Interactive Phenomics Resource: an integrated online phenomics knowledgebase for health data users. Journal of the American Medical Informatics Association, 2024. 31(5): p. 1126-1134. https://doi.org/10.1093/jamia/ocae042

Acknowledgement: The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States Government.

This research used resources at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725 and the Department of Veterans Affairs Office of Information Technology Inter-Agency Agreement with the Department of Energy under IAA No. VA118-16-M-1062.