AI and Aphasia: A Novel Empirical Investigation

Jack Xu, Dr. Naeem Seliya, Molly Heidorn, Dr. Tom Sather

University of Wisconsin, Eau-Claire

Indroduction

Aphasia is an acquired language disorder—often resulting from stroke or brain injury—that impairs a person's ability to speak, understand, read, or write.

Subtyping aphasia is critical because each subtype is correlated with impacts to different regions of the brain, and also each type responds best to different therapeutic approaches.

However, the current gold-standard assessments, including the Western Aphasia Battery (WAB), require around 60–90 minutes of clinician-administered tasks plus manual scoring.

In contrast, our objective is to investigate AI-based solutions to assist SLPs, beginning with an in-depth predictive analysis.

To best represent the semantics of the transcribed text for word embeddings, a novel tokenization of the text was implemented, guided by an expert.

Subsequently, multiple goal-specific predictive analyses are investigated for the five most observed aphasia types.

Some aphasia types are more frequently observed compared to others, adding to the complexity of aphasia-centric data analysis. Ongoing work involves analyzing combinations of class ratios and data sampling and augmentation techniques to address class imbalance.

Objectives

- Preprocess and label data
- Establish baseline performance
- Fine-tune BERT utilizing 3 different classification heads
- Tune key hyperparameters
- Evaluate the models on cross-validated accuracy, macro-F1, AUPRC, AUROC, and G-mean recall
- Test against gradient boosting methods
- Explore data oversampling and undersampling
- Test deep-learning and rule-based data augmentation methods on the transcripts
- Test XGBoost on tuned embeddings

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Google, and A. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019.

[2] J. Wei and K. Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," *arXiv:1901.11196 [cs]*, Aug. 2019, Available: https://arxiv.org/abs/1901.11196

[3] M. Lewis *et al.*, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, doi: https://doi.org/10.18653/v1/2020.acl-main.703.

[4] S. Zhang, L. Jiang, and J. Tan, "Dynamic Nonlinear Mixup with Distance-based Sample Selection," *ACL Anthology*, pp. 3788–3797, Oct. 2022, Accessed: Jul. 21, 2025. [Online]. Available: https://aclanthology.org/2022.coling-1.333/

[5] R. Sennrich, B. Haddow, and A. Birch, "Improving Neural Machine Translation Models with Monolingual Data," *arXiv:1511.06709 [cs]*, Jun. 2016, Accessed: Jan. 25, 2023. [Online]. Available: https://arxiv.org/abs/1511.06709

[6] S. Kobayashi, "Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations," *ACLWeb*, Jun. 01, 2018. https://aclanthology.org/N18-2072 (accessed May 26, 2022).

[7] Adir Rahamim, G. Uziel, E. Goldbraich, and Ateret Anaby Tavor, "Text Augmentation Using Dataset Reconstruction for Low-Resource Classification," Jan. 2023, doi: https://doi.org/10.18653/v1/2023.findings-acl.466.

[8] V. Verma *et al.*, "Manifold Mixup: Better Representations by Interpolating Hidden States," *arXiv (Cornell University)*, Jun. 2018, doi: https://doi.org/10.48550/arxiv.1806.05236.

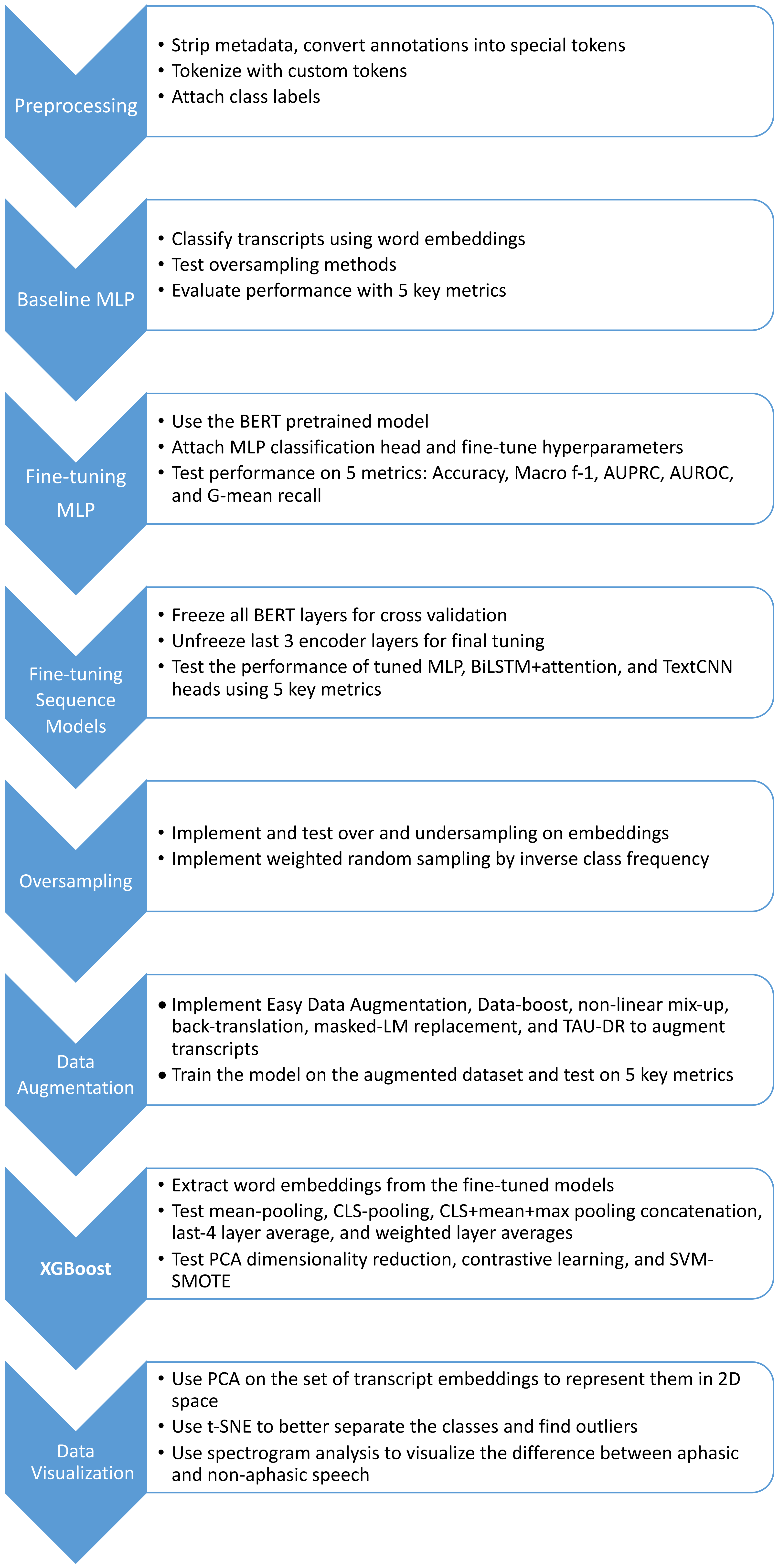
[9] V. Kumar, H. Glaude, de Lichy, and W. Campbell, "A Closer Look At Feature Space Data Augmentation For Few-Shot Intent Classification," *arXiv.org*, 2019. https://arxiv.org/abs/1910.04176

[10] S. Ren, J. Zhang, L. Li, X. Sun, and J. Zhou, "Text AutoAugment: Learning Compositional Augmentation Policy for Text Classification," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Jan. 2021, doi: https://doi.org/10.18653/v1/2021.emnlp-main.711.

Acknowledgements

This research was funded by the National Science Foundation (NSF) Research Experience for Undergraduates (REU) grant OAC-2447779. The computational resources of this study were provided by the Blugold Center for High-Performance Computing under NSF grant CNS-1920220

Process/Methodology



Preprocessing
- Strip metadata, convert annotations into special tokens
- Tokenize with custom tokens
- Attach class labels

Baseline MLP
- Classify transcripts using word embeddings
- Test oversampling methods
- Evaluate performance with 5 key metrics

Fine-tuning MLP
- Use the BERT pretrained model
- Attach MLP classification head and fine-tune hyperparameters
- Test performance on 5 metrics: Accuracy, Macro f-1, AUPRC, AUROC, and G-mean recall

Fine-tuning Sequence Models
- Freeze all BERT layers for cross validation
- Unfreeze last 3 encoder layers for final tuning
- Test the performance of tuned MLP, BiLSTM+attention, and TextCNN heads using 5 key metrics

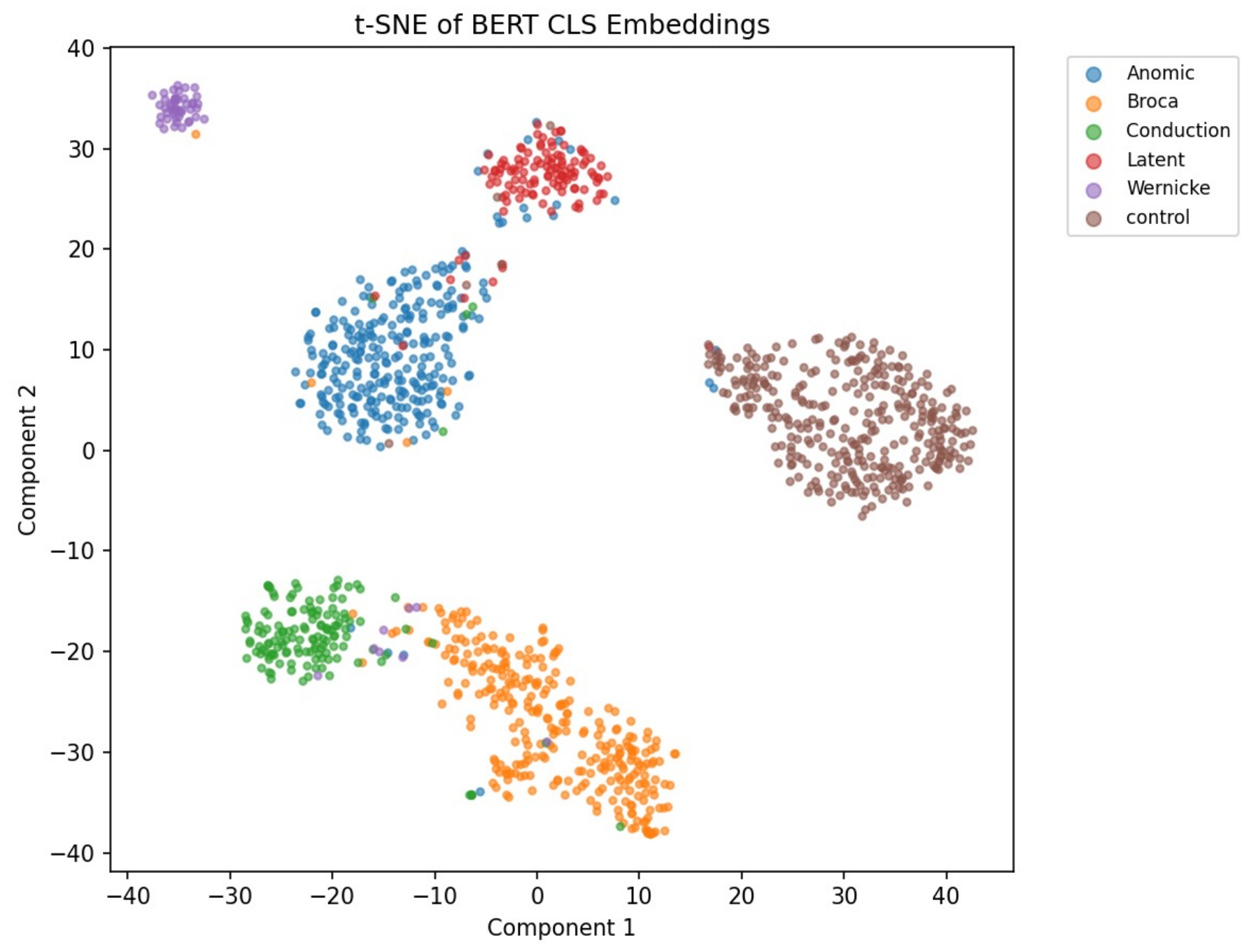
Oversampling
- Implement and test over and undersampling on embeddings
- Implement weighted random sampling by inverse class frequency

Data Augmentation
- Implement Easy Data Augmentation, Data-boost, non-linear mix-up, back-translation, masked-LM replacement, and TAU-DR to augment transcripts
- Train the model on the augmented dataset and test on 5 key metrics

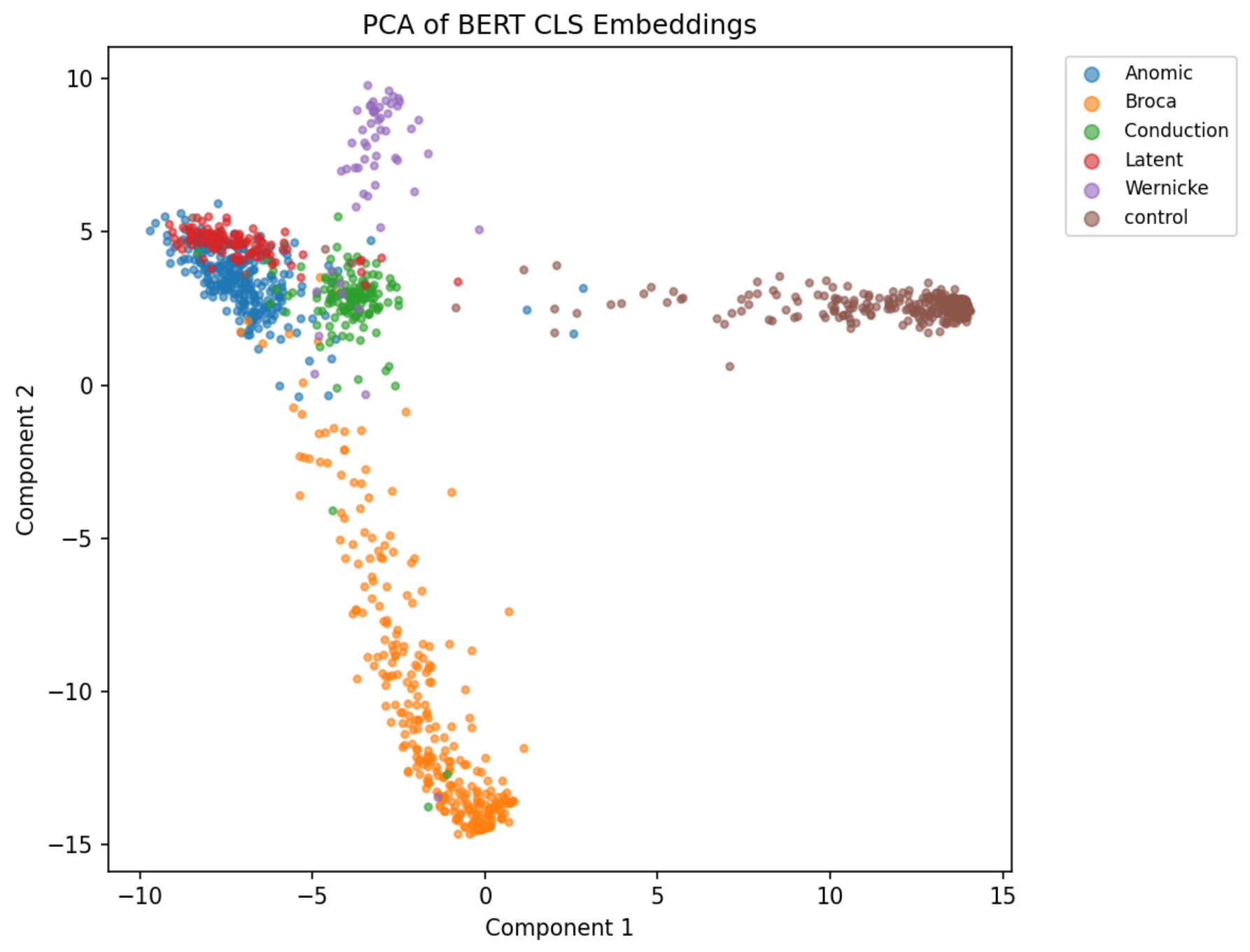
XGBoost
- Extract word embeddings from the fine-tuned models
- Test mean-pooling, CLS-pooling, CLS+mean+max pooling concatenation, last-4 layer average, and weighted layer averages
- Test PCA dimensionality reduction, contrastive learning, and SVM-SMOTE

Data Visualization
- Use PCA on the set of transcript embeddings to represent them in 2D space
- Use t-SNE to better separate the classes and find outliers
- Use spectrogram analysis to visualize the difference between aphasic and non-aphasic speech

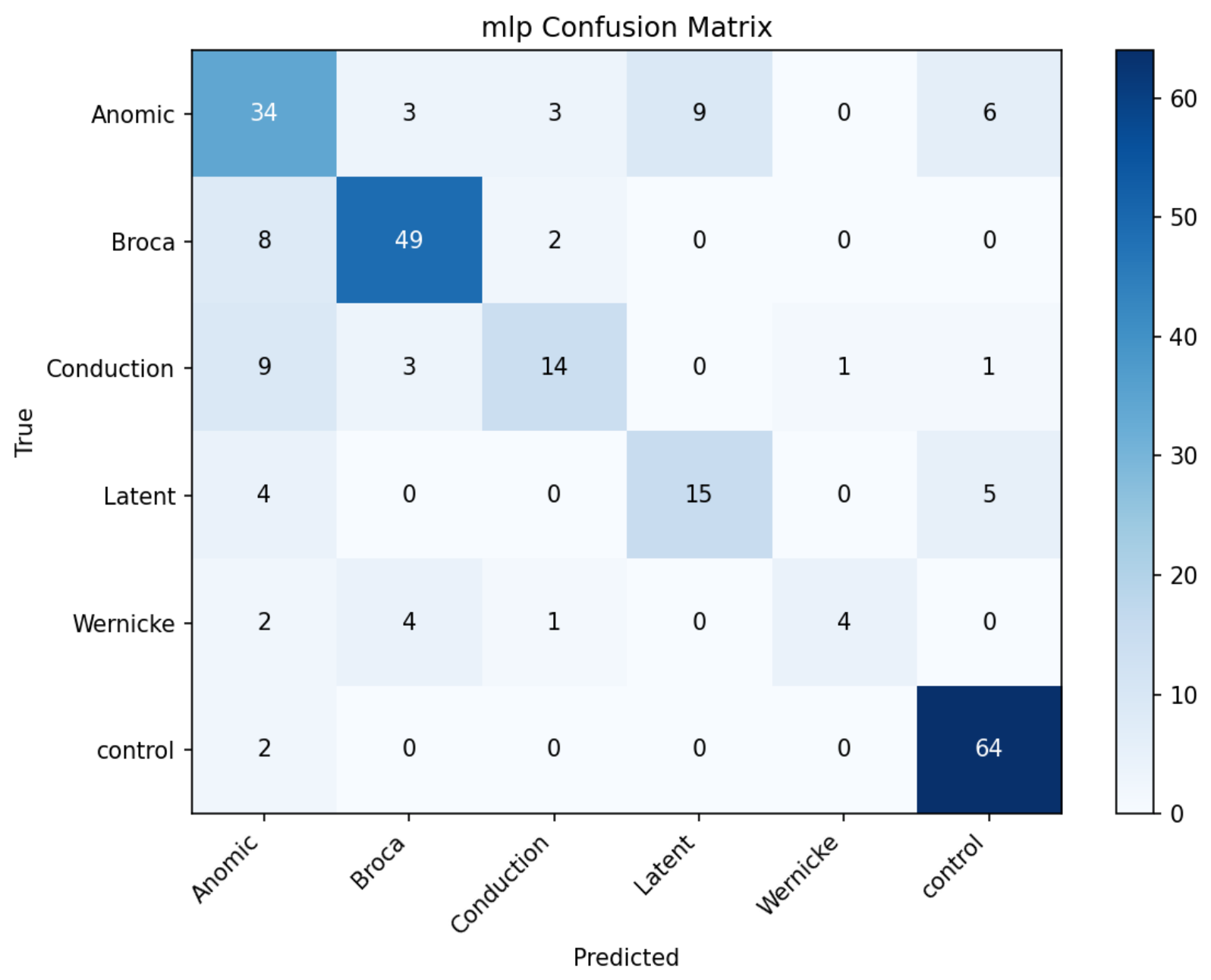
Visualization + Results



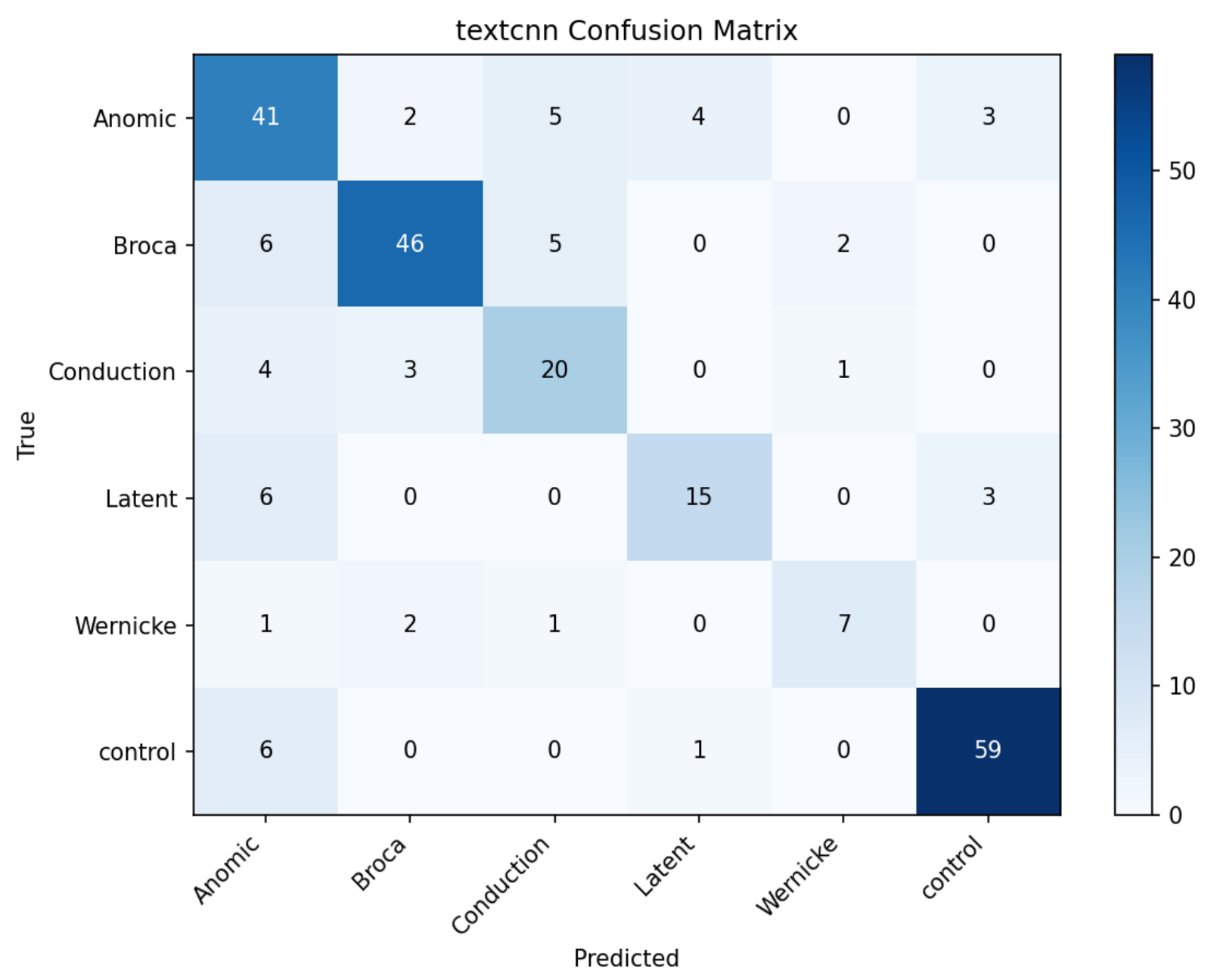
t-SNE of BERT CLS Embeddings



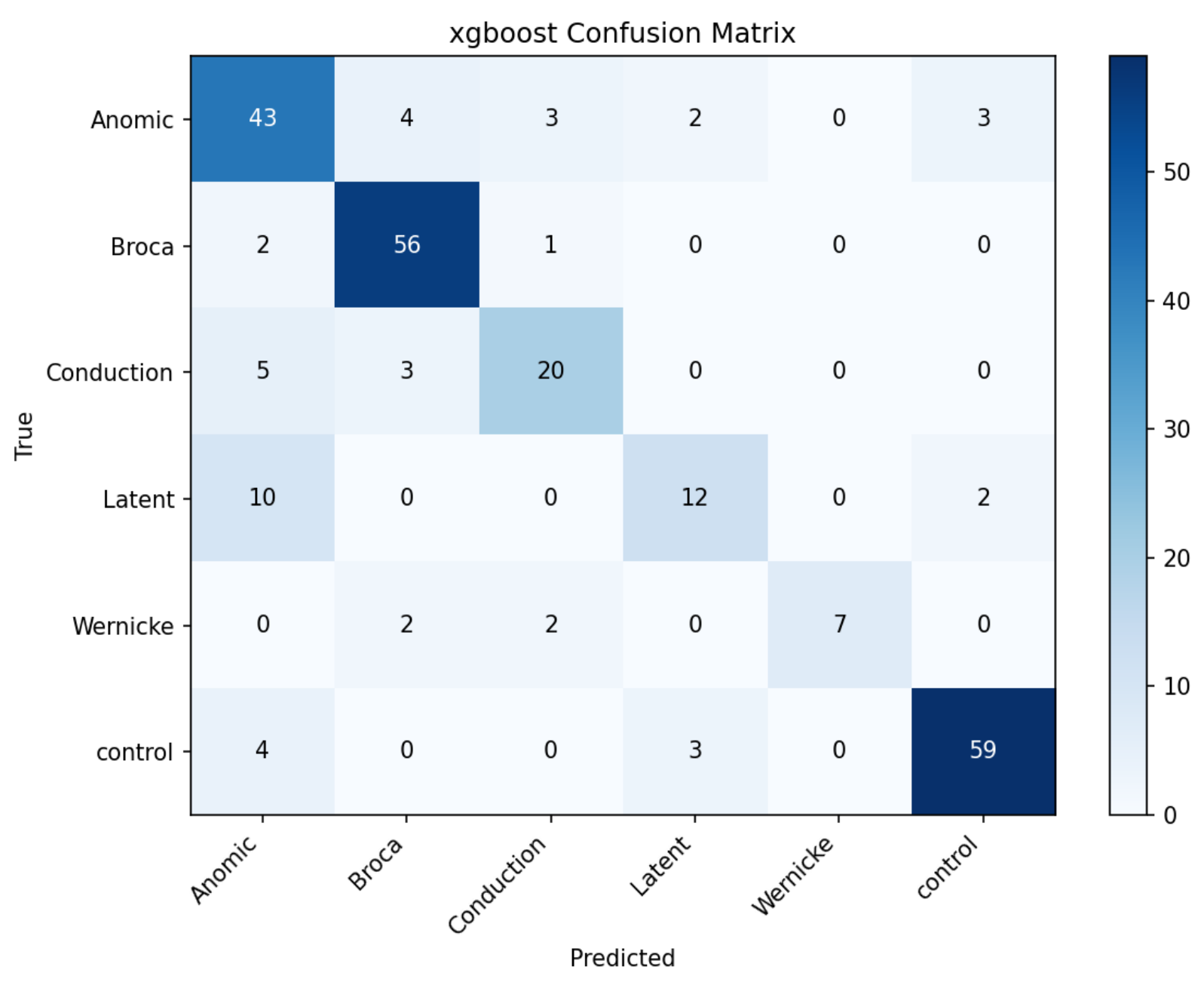
PCA of BERT CLS Embeddings



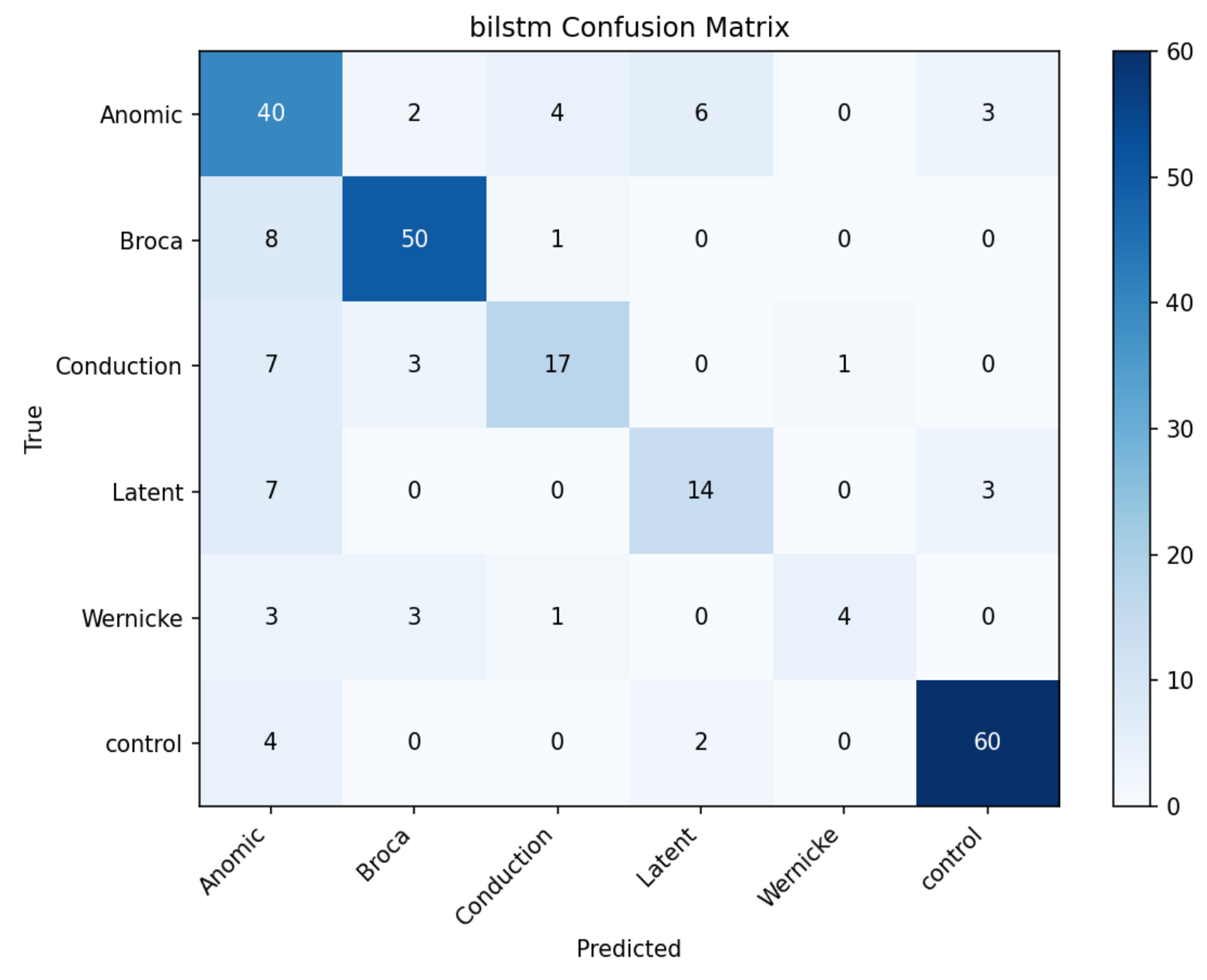
mlp Confusion Matrix



textcnn Confusion Matrix



xgboost Confusion Matrix



bilstm Confusion Matrix

Conclusions

- The MLP baseline on mean-pooled CLS embeddings reached ~67% accuracy (Macro-F1 0.60), performing well on majority but poorly on minority classes.
- Fine-tuning BERT with an MLP head improved to ~74% accuracy (Macro-F1 0.673).
- The BiLSTM+Attention head improved performance (~76% accuracy / 0.697 Macro-F1)
- The TextCNN head delivered the best fine-tuned performance (~77% accuracy / 0.740 Macro-F1), confirming that convolutional filters are strong at capturing local patterns in transcripts.
- The XGBoost model yielded the strongest overall results, achieving an accuracy of 81% and a 0.773 Macro-F1, suggesting that tree-based learners can capitalize on learned representations.

Challenges & limitations

- Rare classes (e.g., Wernicke) still incurred low recall (~0.64) despite weighted sampling; oversampling often hurt due to synthetic samples.
- Transcript-level variability such as length, and disfluencies complicates modeling under fixed maximum sequence lengths.
- Using uniform hyperparameters across augmentation methods may under-utilize each technique's strengths.

Future directions:

- Bootstrap confidence intervals for all metrics to quantify statistical uncertainty.
- Tailor data augmentation per class and head
- Explore hierarchical models that aggregate multiple transcripts or time segments.
- Although other pretrained models yielded similar or worse results in preliminary tests, further tuning of models may uncover gains.