Report: Neurodegenerative Disease

1. Executive Summary

The objective of this assignment was to classify research paper abstracts into five semantically related neurodegenerative categories: Alzheimer’s, Parkinson’s, Vascular Dementia, Mild Cognitive Impairment, and Lewy Body Dementia.

Using a Linear Support Vector Machine as the Champion Model, we achieved an accuracy of 63.33% on the 100-word dataset. We performed a "Harder Prediction" experiment by reducing text length to 20 words, observing an accuracy drop of 8.67% and an AUC drop of 0.0177, confirming the correlation between information density and model performance. Detailed error analysis revealed specific biases against the "Alzheimer's" and "Vascular Dementia" categories, likely due to semantic overlap.

2. Methodology & Data Preparation

2.1 Data Acquisition and Preprocessing

- Source: Europe PMC API.
- Categories: 5 distinct but related topics (Alzheimer, Parkinson, Vascular Dementia, MCI, DLB).
- Partitions: Data was stratified into Training and Testing sets to ensure class balance.
- Preprocessing: Cleaning involved removing stop-words and garbage characters.
- Feature Engineering: Text was transformed using TF-IDF with N-grams (1,2) to capture semantic phrases.
- Dataset Versions:
    1. Standard: ~100 words per document (Title + Abstract).
    2. Hard Mode: Truncated to ~20 words per document to test model robustness.

2.2 Model Selection

We evaluated multiple algorithms including Naive Bayes, Random Forest, KNN, XGBoost, and SVM using 10-Fold Cross-Validation on the training set.

- Champion Model: SVM.
- Reasoning: SVM demonstrated the best balance of Mean Accuracy and low Variability (lowest std dev) during the cross-validation phase.

3. Champion Model Evaluation (100-Word Dataset)

3.1 Performance Metrics

- Accuracy: 63.33%
- Macro Average F1-Score: 0.63
- Weighted Average F1-Score: 0.63

3.2 Confusion Matrix Analysis

Referring to the Confusion Matrix, we observed specific patterns of confusion:

- Strong Performance:
  - MCI: 23 correct predictions (Precision: 0.77).
  - Lewy Body Dementia: 22 correct predictions.
- Areas of Confusion:
  - Vascular Dementia vs. Alzheimer's: 8 instances of Vascular Dementia were misclassified as Alzheimer's. This is semantically explainable as both share overlapping symptoms and "dementia" terminology.
  - Alzheimer's: Had the lowest precision (0.45), often confused with Vascular Dementia. This suggests the model struggles to differentiate the specific "Alzheimer" linguistic markers from general "Dementia" terms in the 100-word context.

4. The Harder Prediction Experiment

To gauge the bias and variability of the model, we constrained the available information by truncating records to the first 20 words.

4.1 Accuracy Drop

- Original Accuracy (100 words): 63.33%
- Harder Accuracy (20 words): 54.67%
- Net Drop: 8.67%

Observation: The 8.67% drop indicates the model is somewhat robust. It suggests that discriminative keywords often appear early in the title or the first sentence of the abstract, allowing the model to maintain decent performance even with limited context.

4.2 AUC Analysis

- Original AUC: 0.8598
- Harder AUC: 0.8421
- Drop: 0.0177

The decrease in AUC confirms that the model's fundamental ability to distinguish between classes (Separability) has degraded with less data, even if the accuracy drop wasn't catastrophic.

5. Bias and Variability Analysis

We overlaid different factors to understand *why* the model fails in the "Harder Prediction" (20-word) scenario.

5.1 Temporal Bias (Date of Publication)

Referring to the Temporal Bias plot:

- 2016-2020: Accuracy was 1.0 (100%).
- Post-2020: Accuracy dropped significantly to ~53.7% (below average).
- Insight: The model exhibits a recency bias. It struggles with the newest papers. This could be due to evolving terminology in the field or a distribution shift in the test set where newer papers might have more abstract titles that require more than 20 words to understand.

5.2 Category Bias

Referring to the Category Bias plot:

- Lowest Accuracy: Alzheimer's (Red bar, well below average).
- Highest Accuracy: MCI and Lewy Body Dementia (Green bars).
- Insight: Reducing the text to 20 words disproportionately hurts the "Alzheimer's" category. This suggests that Alzheimer's papers might require more context to differentiate from general dementia papers, whereas MCI and Lewy Body papers might contain distinct keywords early in the text.

5.3 Factor Analysis

Referring to the SHAP analysis (Image image_2d1dea.png):

- Key Discriminators: The terms "ad" (Alzheimer's Disease) and "pd" (Parkinson's Disease) have high impact.
- Confusion Source: The word "cancer" appears as a feature. This is unexpected in a neurodegenerative dataset and might represent a "garbage" feature or a specific subset of papers discussing co-morbidities, potentially confusing the model. The heavy reliance on acronyms like "ad" explains why the 20-word model remained relatively robust—if the acronym appears early, the model guesses correctly.

Future Work: To improve the Alzheimer's classification, we would recommend increasing the N-gram range to capture specific phrases (e.g., "amyloid plaques") rather than just acronyms, and addressing the temporal bias by re-balancing the training data with more recent examples.