Research/Science Fair

Probing, Patching, and Predicting: Mechanistic Interpretability of Wav2Vec2 Representations for Clinically-Grounded Parkinson’s Disease (PD) Detection

Rationale

Current approaches to Parkinson's disease (PD) voice detection suffer from a critical limitation: they treat neural networks as black boxes. Over 69 published studies use machine learning for PD voice classification, yet none explain how models distinguish pathological from healthy speech. This opacity creates three problems: (1) models fail to generalize across datasets because they may learn spurious correlations (microphone type, recording environment) rather than genuine biomarkers; (2) clinicians cannot verify whether model predictions align with known PD speech pathology (jitter, shimmer, reduced harmonic-to-noise ratio); and (3) regulatory approval requires explainability for medical devices. Post-hoc explainability methods like SHAP and LIME identify which input features correlate with predictions but cannot reveal the internal algorithms the model uses. These methods are fundamentally correlational: they tell us "the model attended to this audio segment" but not "the model detected elevated jitter in this frequency band, which it learned correlates with vocal fold rigidity." This distinction matters: a model could achieve high accuracy by detecting recording quality differences between clinical and home recordings rather than genuine PD biomarkers. Mechanistic interpretability offers a solution by reverse-engineering the internal computations of neural networks. By applying probing classifiers to intermediate representations, we can determine where clinical features (jitter, shimmer, HNR) are encoded. By using activation patching, we can establish which internal features causally affect predictions. This project would be the first to apply mechanistic interpretability to any speech-based disease detection system, addressing a critical gap at the intersection of AI safety and medical AI. The timing is ideal: the paper "Beyond Transcription: Mechanistic Interpretability in ASR" (arXiv:2508.15882, 2025) recently demonstrated that activation patching and probing classifiers can be applied to Whisper and audio transformers. This methodology blueprint can be adapted for PD detection models.

Research Question/Hypothesis(es)/Engineering Goal(s)

Primary Research Question

What internal representations do speech transformers (Wav2Vec2, HuBERT) develop when fine-tuned for Parkinson's disease detection, and which learned features causally determine PD/healthy classification?

Testable Hypotheses

Hypothesis 1 (Layer-wise Clinical Encoding): Clinical voice biomarkers (jitter, shimmer, HNR, F0 variability) are linearly decodable from specific transformer layers, with prosodic features (pitch variation) encoded in middle layers (5-8) and phonatory features (jitter, shimmer) encoded in early layers (2-4).

- **Prediction:** Linear probing classifiers trained on early-layer activations will achieve >80% accuracy predicting binary jitter classification (high/normal), while probes on later layers will show degraded performance.
- **Falsifiable:** If probing accuracy for clinical features is at chance level across all layers, the model does not encode these features in a linearly separable manner.

Hypothesis 2 (Causal Feature Dependency): The model's PD classification depends causally on internal representations that correlate with established clinical biomarkers, not on spurious dataset artifacts.

- **Prediction:** Activation patching experiments will show that perturbing layers/attention heads encoding clinical features (identified via probing) will change PD predictions more than perturbing other components.
- **Falsifiable:** If patching components with high clinical feature probing accuracy does not affect predictions while patching other components does, the model uses features unrelated to known biomarkers.

Hypothesis 3 (Generalization Prediction): Models with internal representations more aligned to clinical biomarkers will generalize better across datasets than models relying on dataset-specific features.

- **Prediction:** A model fine-tuned on PC-GITA (Colombian Spanish) whose probed features align with clinical biomarkers will show <15% accuracy drop on the Italian PVS dataset, compared to >40% drop for a baseline model.

Engineering Goal

Develop an interpretability analysis pipeline that produces clinically meaningful explanations for speech-based PD detection: "This prediction is based on detected elevated jitter (0.024 vs. normal 0.008) in layers 3-4 and reduced HNR (8.2 dB vs. normal 21.4 dB) encoded in attention heads 2.4 and 3.1."

Expected Outcomes

1. **Layer-wise encoding map:** Documentation of which transformer layers encode which clinical features (jitter at layer X, shimmer at layer Y, prosody at layer Z), with probing accuracy curves across all 12 layers.
2. **Causal circuit identification:** Identification of a minimal set of attention heads and MLP layers that are necessary and sufficient for PD classification, validated through activation patching experiments.
3. **Clinical alignment score:** A quantitative metric measuring how much a model's learned features align with established clinical biomarkers (correlation between probed clinical features and model prediction confidence).

4. **Cross-dataset generalization analysis:** Evidence that interpretable models (those using clinically-aligned features) generalize better than black-box models across the PC-GITA, Italian PVS, and mPower datasets.
5. **Interpretable prediction interface:** A prototype that outputs not just "PD probability: 0.87" but "PD indicators: jitter elevated (contribution: 34%), HNR reduced (contribution: 28%), F0 variability reduced (contribution: 21%), other features (17%)."

Materials

- **Primary:** Google Colab Pro ($10/month) with T4/A100 GPU access
- **Backup:** Local computer with 16GB RAM for CPU-based probing analysis
- **Estimated compute:** ~50 GPU-hours total (fine-tuning: 10h, probing: 20h, patching: 20h)

Datasets

Dataset	Size	Language	Access Method
PC-GITA	50 PD + 50 HC, ~6,300 recordings	Spanish	Request from Universidad de Antioquia
Italian PVS	50 participants	Italian	IEEE DataPort (open access)
mPower	>9,500 participants, >65K recordings	English	Synapse Platform (requires data use agreement, syn4993293)
UCI Oxford	31 participants, 195 recordings	English	UCI ML Repository (open access)

Software Libraries

```
# Core ML
torch>=2.0
transformers>=4.35 (Wav2Vec2, HuBERT models)
datasets (HuggingFace)
# Audio Processing
librosa>=0.10
torchaudio>=2.0
parselmouth>=0.4 (Praat interface for jitter/shimmer)
opensmile>=2.4 (acoustic feature extraction)
```

```
# Interpretability

transformer_lens>=1.0 (activation caching patterns)

sklearn (probing classifiers)

numpy, scipy, pandas

# Visualization

matplotlib, seaborn

plotly (interactive figures)
```

Pre-trained Models

- facebook/wav2vec2-base-960h (95M parameters, 12 layers)
- facebook/hubert-base-ls960 (95M parameters, 12 layers)
- facebook/wav2vec2-large-960h (317M parameters, 24 layers) – for validation

Procedures

Phase 1: Dataset Preparation and Baseline Establishment

- **Obtain dataset access:** Submit data use agreements for PC-GITA and mPower. Download Italian PVS and UCI datasets from open sources. Document IRB exemption status (publicly available de-identified data).
- **Standardize audio preprocessing:** Resample all audio to 16kHz mono. Segment sustained vowels (/a/) to 3-second clips with 0.5s overlap. Apply voice activity detection to remove silence. Create train/validation/test splits (70/15/15) stratified by PD/HC label and severity (H&Y stage if available).
- **Extract ground-truth clinical features:** Using Parselmouth (Praat interface), compute for each audio segment:
    - Jitter (local): frequency perturbation quotient
    - Shimmer (local): amplitude perturbation quotient
    - HNR: harmonics-to-noise ratio (dB)
    - F0 statistics: mean, std, range
    - Formant frequencies: F1, F2, F3 means
- Store as metadata for probing experiments.
- **Establish baseline classifier:** Fine-tune wav2vec2-base-960h on PC-GITA for PD classification. Freeze CNN encoder, train only transformer layers + classification head. Use cross-entropy loss, AdamW optimizer (lr=1e-5), 20 epochs with early stopping. Target: >85% test accuracy (comparable to literature baselines).

- **Create minimal pair dataset for patching:** For each PD sample, identify the most acoustically similar HC sample (by MFCC distance). These pairs will be used for activation patching experiments.

Phase 2: Probing Classifier Experiments

- **Extract intermediate representations:** Run all test set samples through fine-tuned model with output_hidden_states=True. Cache hidden states from all 12 transformer layers. Pool across time dimension (mean pooling) to get fixed-length representations per layer.
- **Train probing classifiers for clinical features:** For each layer (0-12) and each clinical feature (jitter_binary, shimmer_binary, HNR_binary, F0_variability_binary):
    - Create binary labels by thresholding at clinical cutoffs (e.g., jitter > 0.02 = abnormal)
    - Train logistic regression probe: probe = LogisticRegression(max_iter=1000)
    - Evaluate with 5-fold cross-validation
    - Record accuracy, F1, and ROC-AUC for each layer-feature combination
- **Create layer-wise encoding heatmap:** Plot 12×4 matrix (layers × clinical features) with probing accuracy as cell values. Identify which layers maximally encode each feature. Statistical test: is peak layer significantly better than chance (permutation test, n=1000)?
- **Validate probing with control tasks:** Repeat probing for control variables that should NOT be predictive of PD:
    - Recording ID (should be chance level if no data leakage)
    - Arbitrary audio segment index
- Selectivity score = (target accuracy - control accuracy) must be >20% for valid features.

Phase 3: Activation Patching Experiments

- **Implement activation patching infrastructure:** Adapt TransformerLens hook patterns for Wav2Vec2/HuBERT:

```
def get_activation_patching_hook(source_activation, position):
    def hook(module, input, output):
        output[:, position, :] = source_activation[:, position, :]
        return output
    return hook
```

- **Layer-level patching:** For each layer (1-12):

    - Run model on HC sample, cache layer activations
    - Run model on matched PD sample with patched HC activations at target layer
    - Measure: Does prediction shift toward HC?
    - Metric: $\Delta \text{logit}(PD) = \text{logit}(PD|\text{patched}) - \text{logit}(PD|\text{original})$
- **Attention head-level patching:** For layers identified as important in step 11:
    - Patch each attention head individually (12 heads × important layers)
    - Identify which specific heads affect PD prediction most
    - Create attention head importance ranking
- **Path patching for clinical features:** Test if heads with high clinical feature probing accuracy are the same heads that causally affect predictions:
    - Hypothesis: If head H encodes jitter (per probing), patching H should change predictions for samples with abnormal jitter
    - Test: Stratify samples by clinical feature values, measure patching effect by stratum
- **Validate with mean ablation:** Complement patching with ablation:
    - Replace target component activations with dataset mean
    - Compare: ablation effect vs. patching effect
    - Concordance validates that components are genuinely necessary, not just involved

Phase 4: Cross-Dataset Generalization Analysis

15. **Train dataset-specific models:** Fine-tune separate Wav2Vec2 models on:
    - PC-GITA only (Model A)
    - Italian PVS only (Model B)
    - mPower only (Model C)
16. **Cross-dataset evaluation matrix:** Test each model on all datasets. Record 3×3 accuracy matrix. Identify which models generalize and which fail.
17. **Compare probing profiles:** For each model, extract layer-wise clinical feature encoding map. Compute "clinical alignment score" = average probing accuracy across clinical features.
18. **Test generalization-interpretability correlation:** Statistical test: Do models with higher clinical alignment scores show smaller cross-dataset accuracy drops? (Spearman correlation between alignment score and generalization gap.)

Phase 5: Synthesis and Validation

- **Build interpretable prediction interface:** Create function that outputs:

  ```
  Input: audio file
  Output: {
    "pd_probability": 0.87,
    "feature_contributions": {
      "jitter_elevated": 0.34,
      "hnr_reduced": 0.28,
      "f0_unstable": 0.21,
      "other": 0.17
    },
    "evidence_layers": [3, 4, 7],
    "key_attention_heads": [(3, 4), (4, 2), (7, 8)]
  }
  ```

- **Clinical validation:** Present 20 predictions with explanations to a clinician (if accessible via mentor network). Ask: "Does this explanation align with known PD speech pathology?" Document clinical face validity.
- **Write research paper:** Structure: Introduction (problem + gap), Methods (probing, patching, metrics), Results (encoding maps, causal circuits, generalization), Discussion (clinical implications, limitations).
- **Prepare science fair materials:** Create poster, display, and demonstration of interpretable prediction interface.

After identifying which layers encode clinical features via probing, train sparse autoencoders on those layers to decompose distributed representations into monosemantic features. This would reveal whether the model has learned abstract 'voice quality' features that combine multiple clinical biomarkers.

Risk Assessment

Risk	Likelihood	Impact	Mitigation
Dataset access delays (PC-GITA, mPower approval)	Medium	High	Start with freely available UCI and Italian PVS datasets. PC-GITA requests typically approved within 2-4 weeks. Begin probing experiments on UCI while waiting.
Low probing accuracy (clinical features not linearly encoded)	Medium	Medium	This is a valid negative result. Pivot to nonlinear probes (small MLP) or SAE-based feature discovery. The finding that clinical features are not linearly encoded is itself publishable.
Compute limitations (Colab GPU quotas)	Medium	Medium	Use Colab Pro ($10/mo). Cache all intermediate activations to disk after one forward pass. Probing and analysis can run on CPU. Limit patching experiments to most promising layers.
Model fails to achieve baseline accuracy	Low	High	Use established hyperparameters from PD voice detection literature. Try both Wav2Vec2 and HuBERT. If both fail, use pre-trained models without fine-tuning (transfer learning probe).
No causal relationship found (patching doesn't affect predictions)	Medium	Medium	Valid negative result indicating model uses distributed representations. Report this and pivot to SAE-based feature discovery for decomposing distributed features.
Clinical validation not possible (no clinician access)	Medium	Low	Focus on objective metrics (probing accuracy, cross-dataset generalization). Clinical validation is valuable but not required for ISEF.

Safety and Ethical Considerations

- All datasets are publicly available or available under standard research data use agreements
- No collection of new human subjects data; no IRB required
- Project does not develop diagnostic tools for clinical use; explicitly framed as interpretability research
- Published results will include limitations about applicability to clinical settings

Data Analysis

Statistical Methods

Probing accuracy analysis:

- Report mean ± standard deviation across 5-fold CV
- Compare layers using paired t-tests with Bonferroni correction
- Effect size: Cohen's d for layer differences
- Visualization: Line plots with 95% confidence intervals

Activation patching analysis:

- Metric: Δ logit(PD) = change in PD log-odds after patching
- Statistical test: Permutation test ($n=1000$) for significance of patching effect
- Baseline: Mean Δ logit across random patches (null distribution)
- Threshold for importance: Patching effect > 2 standard deviations above baseline

Cross-dataset generalization:

- Primary metric: Absolute accuracy drop (train dataset → test dataset)
- Secondary metric: Relative accuracy drop (% of original performance retained)
- Correlation: Spearman ρ between clinical alignment score and generalization

Quantitative success criteria:

1. At least one clinical feature shows significantly above-chance probing accuracy (>65%) at some layer
2. Activation patching identifies specific layers/heads that change predictions by >0.5 logits
3. Clinical alignment score correlates with generalization ($\rho > 0.4$, $p < 0.05$)

Visualization Plan

- **Figure 1:** Layer-wise probing accuracy heatmap (layers × clinical features)
- **Figure 2:** Activation patching effect by layer (bar chart with error bars)

- **Figure 3:** Attention head importance matrix for top 3 layers
- **Figure 4:** Cross-dataset accuracy matrix (3×3 heatmap)
- **Figure 5:** Scatter plot of clinical alignment vs. generalization gap
- **Figure 6:** Example interpretable prediction with explanation breakdown

Bibliography

Foundational Mechanistic Interpretability

1. Elhage, N., et al. (2021). "A Mathematical Framework for Transformer Circuits." *Transformer Circuits Thread*, Anthropic. https://transformer-circuits.pub/2021/framework/
2. Elhage, N., et al. (2022). "Toy Models of Superposition." *Transformer Circuits Thread*, Anthropic. https://transformer-circuits.pub/2022/toy_model/
3. Bricken, T., et al. (2023). "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning." *Transformer Circuits Thread*, Anthropic. https://transformer-circuits.pub/2023/monosemantic-features/
4. Heimersheim, S., & Nanda, N. (2024). "How to use and interpret activation patching." *arXiv:2404.15255*.
5. Bereska, L., & Gavves, E. (2024). "Mechanistic Interpretability for AI Safety — A Review." *arXiv:2404.14082*.

Speech Model Interpretability

6. Pasad, A., et al. (2023). "What Do Self-Supervised Speech Models Know About Words?" *Transactions of the Association for Computational Linguistics*, 12, 372-391.
7. Jin, Z., et al. (2025). "Beyond Transcription: Mechanistic Interpretability in Automatic Speech Recognition." *arXiv:2508.15882*.
8. Mohamed, A., et al. (2022). "Self-Supervised Speech Representation Learning: A Review." *IEEE Journal of Selected Topics in Signal Processing*, 16(6), 1179-1210.
9. Chen, Y., et al. (2025). "A Large-Scale Probing Analysis of Speaker-Specific Attributes in Self-Supervised Speech Models." *arXiv:2501.05310*.

Parkinson's Voice Detection

10. Moro-Velazquez, L., et al. (2023). "On inter-dataset generalization of machine learning approaches to Parkinson's disease detection from voice." *Computer Speech & Language*, 82, 101535.

11. Bot, B.M., et al. (2016). "The mPower study, Parkinson disease mobile data collected using ResearchKit." *Scientific Data*, 3, 160011.
12. Orozco-Arroyave, J.R., et al. (2014). "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease." *LREC 2014*.
13. Ramezani, A., et al. (2025). "Voice-Based Detection of Parkinson's Disease through Machine Learning and Deep Learning Methods: A Systematic Review." *MDPI Electronics*.

Concept Bottleneck Models

14. Koh, P.W., et al. (2020). "Concept Bottleneck Models." *ICML 2020*.
15. Oikarinen, T., et al. (2023). "Label-free Concept Bottleneck Models." *ICLR 2023*.

ISEF Category Recommendation

Computational Biology and Bioinformatics or **Systems Software**

What Makes This ISEF Grand Prize Caliber

1. **Novel intersection:** First application of mechanistic interpretability to any speech-based disease detection. Combines cutting-edge AI safety research (Anthropic's methods) with high-impact medical AI.
2. **Technical sophistication:** Uses advanced methods (probing classifiers, activation patching, causal analysis) that are active areas of research at leading AI labs. Not just "apply pre-trained model."
3. **Real-world impact:** Addresses FDA-recognized need for explainable medical AI. Could influence how speech biomarker models are developed and validated.
4. **Generalizable methodology:** Pipeline can be applied to other speech pathology detection (ALS, stroke, depression) and other modalities (ECG, imaging).
5. **Quantifiable results:** Clear metrics (probing accuracy, patching effects, generalization scores) enable rigorous evaluation of claims.

Mentorship Recommendations

- University speech pathology department (clinical feature validation)
- Computer science professor with ML interpretability focus
- Contact researchers from "Beyond Transcription" paper for methodology guidance