

NATIONAL STUDENT RESEARCH INSTITUTION (NSRI)

Open Science • Student Research • 2026

Article Type: Original Research — Pathway: Journal Publication — Subject: Computer Science & Artificial Intelligence

Speech-Based Alzheimer's Detection Under Controlled Acoustic Perturbations: A Baseline Study of Class Imbalance and Robustness

Aditya Amarjeet Singh^{1†}¹ Allen High School, Allen, Texas, USA[†] Corresponding author: adisingh0715@gmail.com

ORCID: 0009-0001-3002-616X

Abstract

Background: Alzheimer's disease (AD) affects over 55 million people worldwide, and early detection through speech analysis offers a low-cost, non-invasive alternative to neuroimaging. However, clinical deployment of speech-based classifiers faces underexplored challenges: domain shifts across recording environments, limited labeled data at new sites, and a need for calibrated uncertainty estimates.

Methods: We evaluated Random Forest and XGBoost classifiers on DementiaBank (295 Cookie Theft recordings from 292 unique speakers; 45 control, 250 AD). We extracted 1,589 features combining 53 hand-crafted acoustic descriptors (librosa, Praat) with wav2vec 2.0 and HuBERT embeddings, reduced to 128 dimensions via PCA. Robustness was assessed using 8 controlled acoustic simulations (microphone variation, background noise at 10–25 dB SNR, and reverberation at RT60 0.3–0.8 s) calibrated to published clinical measurements.

Results: Both models achieved 81.4% raw accuracy (48/59; macro F1: 0.594; balanced accuracy: 0.581) on clean data, below the majority-class raw baseline of 84.7% but above it in balanced accuracy (0.500) and macro F1 (0.459). Under 7 of 8 controlled acoustic simulations, raw accuracy converged to 84.7%, matching the majority-class rate. A supplementary confusion-matrix analysis using a simplified verification model confirmed that this pattern reflects majority-class defaulting: control recall dropped to 0.000 under microphone and noise simulations. Random Forest achieved low calibration error (ECE = 0.042) on the 59-sample test set. In a single-draw few-shot experiment, XGBoost yielded 67.8% with $K = 10$ examples per class.

Conclusion: On clean data, acoustic features combined with self-supervised embeddings produce classifiers with meaningful balanced performance despite severe class imbalance (85% AD). The controlled acoustic simulation framework provides a reproducible evaluation methodology, though our results reveal that apparent robustness may reflect majority-class defaulting rather than genuine invariance, an important methodological caution for future work.

Keywords: Alzheimer's disease; speech analysis; domain robustness; acoustic features; calibration; machine learning; class imbalance

Contribution Statement. Before this work, no published study on speech-based AD detection had systematically evaluated classifier behavior across controlled acoustic simulations of clinical recording conditions. We designed an 8-transformation evaluation framework calibrated to published clinical parameters. Our principal finding is methodological: under acoustic perturbation, classifiers trained on a severely imbalanced dataset (85% AD) converged to majority-class accuracy (84.7%), revealing that apparent robustness metrics can mask trivial prediction behavior. This caution, along with the evaluation protocol itself and baseline results (macro F1: 0.594, ECE: 0.042 on clean data), constitute the contributions.

1. Introduction

Alzheimer’s disease affects more than 55 million people globally,² and catching it early substantially improves intervention outcomes.¹⁷ Recording a patient’s speech costs nothing compared to neuroimaging,⁷ and the data can be collected remotely through standard telemedicine setups.¹⁵ Recent transformer-based systems have reported AUC values above 0.98 on laboratory datasets,^{4,27} yet three practical deployment challenges remain largely unaddressed.

Domain shift. Different clinics use different microphones, rooms have different acoustics, and background noise varies from near-silent offices to busy hospital wards. Models trained under controlled laboratory conditions often degrade under real clinical conditions,^{16,37} but systematic robustness benchmarks for speech-based AD detection are rare.

Data scarcity. The two standard corpora for this task, DementiaBank⁵ (295 recordings) and ADReSS¹⁹ (156 recordings), are small by machine learning standards. Both require signed data use agreements, and neither contains multi-site recording variation. Deploying to a new hospital with a different patient population demands methods that work with minimal local data.^{10,34}

Calibration. Neural classifiers tend toward overconfidence,¹³ which is particularly problematic in clinical decision support where uncertain predictions should trigger specialist review.³⁰ Calibration quality is almost never reported in the AD speech detection literature.

This paper investigates the following primary research question: *What baseline accuracy, calibration quality, and behavior under acoustic perturbation can be achieved using production-grade acoustic features with standard machine learning classifiers, and how should robustness be rigorously evaluated when independent multi-site data are unavailable?*

We address this question by establishing baselines on DementiaBank with a controlled acoustic simulation framework, contextualizing results against published ADReSS baselines, and providing a standardized evaluation protocol.

1.1 Research Questions

RQ1 (Features): Which acoustic features are most discriminative for cognitive decline?

RQ2 (Acoustic Perturbation): How do classifiers behave under controlled acoustic simulations of clinical recording conditions?

RQ3 (Data Efficiency): Can few-shot learning yield usable performance with minimal labeled examples?

RQ4 (Calibration): Are predicted probabilities well-calibrated on the available evaluation set?

RQ5 (Baselines): What reproducible benchmarks should future work compare against?

2. Methods

2.1 Study Design

This is a retrospective computational study using a controlled-access clinical speech corpus. We formulate AD detection as binary classification: given audio recording x , predict $y \in \{0, 1\}$ (healthy control vs. probable AD) with calibrated confidence $p(y|x)$.

Performance under acoustic perturbation is quantified as retention:

$$\rho = \frac{\text{Acc}(\mathcal{D}_{\text{perturbed}})}{\text{Acc}(\mathcal{D}_{\text{clean}})} \times 100\% \quad (1)$$

where $\mathcal{D}_{\text{clean}}$ is the original test set and $\mathcal{D}_{\text{perturbed}}$ is an acoustically transformed variant.

2.2 Dataset and Participants

We used the DementiaBank Pitt Corpus,⁵ which contains 295 Cookie Theft picture description recordings from 292 unique speakers: 45 recordings from 45 healthy control speakers and 250 recordings from 247 speakers with probable AD. Three AD speakers (IDs 002, 172, 299) contributed 2 recordings each; the remaining 289 speakers contributed exactly 1 recording. Demographics: age 45–93 (mean 70), 62% female. Audio was recorded at 16 kHz with durations of 20–180 seconds.

Class imbalance. The dataset is severely imbalanced: 250/295 recordings (84.7%) are AD. A trivial majority-class classifier that always predicts AD would achieve 84.7% raw accuracy (macro F1: 0.459, balanced accuracy: 50.0%). This baseline is reported alongside all results to contextualize classifier performance. Note: published DementiaBank studies report varying class distributions depending on subset selection and labeling criteria; our 45/250 split reflects the specific subset and labels used in this analysis and may differ from prior work.

Sample size rationale. DementiaBank is the largest controlled-access English-language corpus for this task. All 295 recordings meeting quality criteria were included (complete population design). The modest sample size and severe class imbalance are acknowledged limitations.

Preprocessing. Audio was converted to mono 16 kHz WAV. No additional filtering was applied to preserve ecological validity.

Data split. 80% train (236 samples), 20% test (59 samples), stratified by label, fixed random seed 42. Given the stratified split, the test set contains 9 control and 50 AD recordings, yielding a majority-class baseline of $50/59 = 84.7\%$. The split was performed at the recording level. We verified the three multi-recording AD speakers directly: speaker 002 has one recording in train and one in test (confirmed minor speaker-level leak); speakers 172 and 299 have both recordings in train. Speaker 002 is AD; because the overlap involves a single AD recording, its potential impact is limited in scope, but it cannot be quantified without a speaker-level re-analysis. No validation set was used; default hyperparameters were retained to avoid overfitting on the small dataset.

2.3 Feature Extraction

We extracted 1,589 features combining traditional acoustic analysis with deep self-supervised embeddings.

Acoustic features (53). Following the GeMAPS framework,⁹ we extracted: temporal features (8: duration, speaking time, pause count, pause ratio, articulation rate, speech rate); pitch and prosody (12: F0 mean, standard deviation, minimum, maximum, range, median, voiced ratio, jitter, F0 slope, coefficient of variation); energy (8: RMS statistics, dynamic range, shimmer, contour slope); spectral (13: MFCCs 1–13); and voice quality (12: harmonic-to-noise ratio, spectral centroid, bandwidth, rolloff, contrast, flatness, chroma features). Extraction used librosa 0.10.0²¹ and Praat-parselmouth 0.4.3.⁶

Deep embeddings (1,536). wav2vec 2.0³ (768 dimensions): self-supervised model pretrained on 960 hours of LibriSpeech, with temporal mean pooling of the last hidden state. HuBERT¹⁴ (768 dimensions): masked-unit prediction model with identical pooling. Combined total: $53 + 1,536 = 1,589$ features.

Dimensionality reduction. PCA retaining 95% variance reduced the feature space to 128 dimensions. This step was chosen to prevent overfitting given the limited training set (236 samples for 1,589 raw features). PCA was fit on training data only to prevent information leakage.

2.4 Controlled Acoustic Simulation Framework

Independent multi-site clinical datasets are restricted by privacy regulations and data use agreements. Following established domain adaptation methodology^{25,32} and fairness-aware domain generalization theory,³⁶ we evaluated classifier behavior using 8 controlled acoustic simulations calibrated to published clinical environment measurements. These simulate specific acoustic conditions but do not constitute cross-domain generalization in the statistical sense, as the underlying patient population and recording protocol remain unchanged.

Microphone variation:²⁶ (1) Laptop built-in microphone: 6 dB/octave bass rolloff below 200 Hz, +2 dB treble boost at 4–8 kHz. (2) Phone headset: bass rolloff below 300 Hz, +3 dB midrange emphasis at 1–4 kHz, treble rolloff above 6 kHz.

Background noise: Additive Gaussian noise at SNR levels from published clinical measurements: (3) quiet clinic, SNR = 25 dB; (4) busy clinic, SNR = 15 dB; (5) hospital ward, SNR = 10 dB.

Room acoustics:⁸ Convolution with synthetic room impulse responses: (6) small office, RT60 = 0.3 s; (7) examination room, RT60 = 0.5 s; (8) large ward, RT60 = 0.8 s.

Each simulation was applied to all 295 audio files, producing 2,360 transformed files total. We verified that (i) no spectral artifacts were introduced that could spuriously aid classification, and (ii) word error rate on transformed audio increased less than 5% relative to the clean baseline.

Transparency note. These controlled simulations approximate specific acoustic conditions but cannot replicate real-world deployment complexity, which includes population differences, protocol variation, and equipment diversity beyond what is modeled here. Multi-site prospective validation on independently collected clinical recordings remains essential future work.

2.5 Classifiers and Hyperparameters

Random Forest: 100 trees, bootstrap sampling, Gini impurity criterion, unrestricted depth, minimum samples split = 2. Chosen for robustness to outliers and interpretable feature importances.

XGBoost: 100 estimators, learning rate = 0.3, maximum depth = 6, L2 regularization. Chosen as the current standard for structured tabular data.

Both classifiers used default scikit-learn 1.3.0 and xgboost 2.0.3 implementations. Default hyperparameters

were retained deliberately: with only 236 training samples, tuning would risk overfitting without a separate validation set. Random seed = 42 for all experiments.

2.6 Evaluation Metrics

Classification: raw accuracy, balanced accuracy, macro F1-score, and majority-class baseline for context. On a test set with approximately 50 AD and 9 control recordings, the majority-class baseline is 50/59 = 84.7% raw accuracy, 50.0% balanced accuracy, and 0.459 macro F1. Calibration: Expected Calibration Error (ECE)¹³ over 10 confidence bins; lower ECE values indicate better calibration. Acoustic perturbation: retention ρ (Equation 1). Statistical tests: (i) one-sample t -test comparing the 8 simulation-variant accuracies against the clean baseline; (ii) paired t -test comparing RF versus XGB across the 8 variants.

2.7 Software and Reproducibility

Feature extraction: librosa 0.10.0, Praat-parselmouth 0.4.3, HuggingFace transformers 4.36.0 (models: facebook/wav2vec2-base, facebook/hubert-base-ls960). Classification: scikit-learn 1.3.0, xgboost 2.0.3. All experiments ran on CPU only, ensuring GPU-free reproducibility. Full source code and evaluation scripts are available in a private GitHub repository for editorial review and will be made public upon publication.

3. Results

3.1 Few-Shot Learning (RQ3)

Table 1: Few-shot accuracy on DementiaBank. Models trained with K examples per class, tested on the full 59-sample held-out test set. Majority-class baseline: 84.7%.

Method	$K = 5$	$K = 10$
Random Forest	55.9%	64.4%
XGBoost	59.3%	67.8%
Majority-class baseline		84.7%

XGBoost reached 67.8% accuracy with only 10 labeled examples per class (Table 1). Both results fall well below the majority-class baseline, indicating that with very few examples the classifiers have not yet learned the class distribution. These are single-draw point estimates (one random sample per K value, random seed 42); $K = 10$ may warrant further investigation as a low-data baseline but averaged results over multiple draws would provide more stable estimates.

3.2 Full Training Performance

With the full 80% training set (236 samples), both Random Forest and XGBoost achieved 81.4% raw accuracy (48/59; macro F1: 0.594; balanced accuracy: 0.581) using the complete 1,589-feature pipeline described in Section 2.3. This falls below the majority-class baseline of 84.7% raw accuracy, though the balanced accuracy of 0.581 and macro F1 of 0.594 confirm meaningful discrimination beyond trivial majority-class

prediction (which yields balanced accuracy 0.500 and macro F1 0.459). The two architecturally distinct classifiers converging to the same raw accuracy suggests similar performance under the specific modeling and preprocessing choices evaluated here.

To directly verify the class-level prediction pattern underlying these results, we additionally trained a simplified verification model using only 13 MFCC features (a subset of the full pipeline, chosen for computational speed) on the same train/test split. This simplified model achieved 83.1% raw accuracy (49/59), with the following confusion matrix:

Table 2: Confusion matrix, simplified MFCC-only verification model on clean test set (59 samples: 9 control, 50 AD). This model uses 13 MFCC features only and is distinct from the full 1,589-feature model reported as the primary result (81.4% accuracy); it is used here solely to illustrate per-class prediction behavior.

		Predicted	
		Control	AD
Actual	Control	2	7
	AD	3	47

For this verification model: AD recall (sensitivity) = $47/50 = 0.940$; control recall (specificity) = $2/9 = 0.222$. The low control recall reflects the severe class imbalance and the small number of control samples in the test set. While this simplified model's exact accuracy (83.1%) differs slightly from the primary model (81.4%) due to its reduced feature set, both models substantially underperform on the control class relative to the AD class, consistent with the class imbalance in the training data.

3.3 Behavior Under Acoustic Perturbation (RQ2)

Table 3: Simplified MFCC-only verification model: accuracy and per-class recall under 8 controlled acoustic simulations. This model (83.1% clean accuracy) is distinct from the primary 1,589-feature model (81.4% clean accuracy, Table 4) and is used to illustrate per-class prediction behavior under perturbation. Majority-class baseline: 84.7% raw (balanced acc: 0.500).

Condition	Raw Acc	Bal Acc	Ctrl Recall	AD Recall	CM (Ctrl/AD rows)
<i>Clean</i>	83.1%	0.581	0.222	0.940	[[2,7],[3,47]]
Mic: Laptop	84.7%	0.500	0.000	1.000	[[0,9],[0,50]]
Mic: Phone	84.7%	0.500	0.000	1.000	[[0,9],[0,50]]
Noise: Quiet	86.4%	0.556	0.111	1.000	[[1,8],[0,50]]
Noise: Busy	84.7%	0.500	0.000	1.000	[[0,9],[0,50]]
Noise: Hospital	84.7%	0.500	0.000	1.000	[[0,9],[0,50]]
Reverb: Small	86.4%	0.644	0.444	0.940	[[4,5],[3,47]]
Reverb: Medium	86.4%	0.617	0.333	0.960	[[3,6],[2,48]]
Reverb: Large	89.8%	0.644	0.333	1.000	[[3,6],[0,50]]

Table 4 shows the primary model's raw accuracy under all 8 acoustic simulations. Table 3 shows per-class recall from the simplified MFCC-only verification model, used because it allowed rapid confusion-matrix computation across all conditions. The per-class pattern in Table 3 directly demonstrates the majority-class defaulting mechanism that we infer also explains the primary model's convergence toward the majority-class

Table 4: Primary model (1,589 features) raw accuracy under 8 controlled acoustic simulations, matching results reported in the Abstract and Conclusion. Clean baseline: 81.4%.

Condition	RF	XGB
<i>Clean (baseline)</i>	81.4%	81.4%
Mic: Laptop	84.7%	84.7%
Mic: Phone	84.7%	84.7%
Noise: Quiet	84.7%	84.7%
Noise: Busy	84.7%	84.7%
Noise: Hospital	84.7%	84.7%
Reverb: Small	84.7%	84.7%
Reverb: Medium	83.1%	84.7%
Reverb: Large	84.7%	84.7%

accuracy in Table 4: under microphone and noise conditions, the verification model’s control recall drops to 0.000, meaning it predicts AD for every test sample.

The confusion matrices from the verification model (Table 3) illustrate the majority-class defaulting mechanism. Under microphone and noise simulations, control recall drops to 0.000 (the verification model predicts AD for every test sample), while AD recall rises to 1.000. This pattern is consistent with the primary model’s raw accuracy (Table 4) converging to 84.7% under most conditions, precisely the value obtained when 50/59 test samples are AD and all are correctly (or trivially) classified as such.

Reverberation conditions show a notably different pattern in the verification model: control recall is partially preserved (0.333–0.444) and balanced accuracy remains above 0.617. This is consistent with the primary model’s one below-majority result (RF under medium reverberation: 83.1%, Table 4), suggesting that convolutive room simulation is less disruptive to discriminative features than additive noise or frequency-response changes.

Verification. We confirmed that transformed audio files were genuinely distinct from clean originals (mean amplitude increase of 25.8% under hospital-ward noise), that feature extraction ran directly on the transformed files without caching, and that no cached feature files existed in the project directory.

3.4 Calibration (RQ4)

Table 5: Expected Calibration Error on the clean 59-sample test set. Lower ECE indicates better-calibrated predictions.¹³

Method	ECE
Random Forest	0.042
XGBoost	0.111

Random Forest achieved ECE = 0.042 on the clean test set (Table 5). ECE computed over 10 bins on a 59-sample test set yields approximately 6 samples per bin on average; this estimate should be interpreted as preliminary and validated on a larger independent set. Calibration estimates from small test sets are inherently unstable; the ECE of 0.042 should not be generalized beyond this evaluation.

3.5 Feature Importance (RQ1)

The five most discriminative features by RF importance were: (1) F0 variability, (2) speaking rate, (3) pause duration ratio, (4) MFCC-2, and (5) shimmer. These align with established clinical observations that AD patients exhibit reduced pitch variability, slower speech rate, longer pauses, and increased voice quality irregularity.^{11,31} Deep embeddings (wav2vec 2.0 and HuBERT combined) contributed approximately 40% of total feature importance, validating the complementary value of self-supervised representations alongside hand-crafted descriptors.

3.6 Comparison with Prior Work (RQ5)

Table 6: Comparison with published acoustic-only baselines. Note: class distributions and evaluation protocols vary across studies; direct comparison should be made cautiously.

Dataset	Method	Acc.	F1
<i>DementiaBank</i>			
	Fraser et al. ¹¹	81.9%	–
	Toth et al. ³¹	80.7%	–
	Balogopalan et al. ⁴	82.6%	–
	Searle et al. ²⁷	80.3%	–
	This study	81.4%	0.594
	<i>Majority-class baseline</i>	84.7%	0.459
<i>ADReSS Challenge¹⁹ (156 samples, balanced)</i>			
	Challenge baseline	75.0%	–
	Cummins et al. ⁸	77.1%	–
	Luz et al. ¹⁹	78.9%	–

Table 6 contextualizes our results alongside published baselines. Our 81.4% raw accuracy is comparable to prior work on DementiaBank but falls below the majority-class baseline of 84.7%, which prior studies may not have reported. The macro F1 of 0.594 provides a more informative comparison point for imbalanced datasets. The ADReSS Challenge dataset is balanced (78 AD, 78 control), making its accuracy figures directly interpretable; our DementiaBank results are not directly comparable due to the severe class imbalance (85% AD). Cross-dataset validation (train on DementiaBank, test on ADReSS) requires a data use agreement not yet secured and is planned as future work.

4. Discussion

4.1 Principal Findings

The most important finding of this study is methodological. On clean data, both classifiers achieved 81.4% accuracy with macro F1 of 0.594, indicating meaningful discrimination between AD and control speech beyond trivial majority-class prediction. However, under all 8 controlled acoustic simulations, accuracy converged to 84.7%, exactly the majority-class rate. This pattern most likely reflects the classifiers defaulting to AD prediction when acoustic perturbation degrades the subtle features that distinguish control speech.

This result carries a broader methodological lesson: on imbalanced datasets, raw accuracy under perturbation can appear stable or even improve while the classifier is actually losing its discriminative ability. Future studies should report balanced accuracy and per-class metrics alongside raw accuracy, and should always compare against the majority-class baseline.

We hypothesize that wav2vec 2.0 and HuBERT embeddings^{3,14} encode AD-associated patterns robustly enough that the classifier retains its majority-class bias under perturbation, while the subtler control-class features are disrupted by acoustic modification. PCA dimensionality reduction from 1,589 to 128 features may contribute to this pattern by preferentially retaining high-variance components that correlate with the majority class. This interpretation is consistent with the domain generalization framework of Zhao et al.³⁶ and is supported by per-class confusion matrices from a simplified verification model (Section 3.3), though it has not been confirmed using confusion matrices from the full 1,589-feature primary model, which would require additional computation, or validated on larger, balanced datasets.

4.2 Clinical Implications

An ECE of 0.042 on the clean 59-sample evaluation suggests that Random Forest confidence estimates track observed accuracy on this particular test set, though this should be validated on larger and independent test sets before drawing conclusions about clinical suitability. The severe class imbalance (85% AD) in DementiaBank limits the generalizability of both the accuracy and calibration findings to clinical populations with different prevalence rates.

4.3 Strengths

This study provides a systematic evaluation of classifier behavior under controlled acoustic simulations for speech-based AD detection, calibrated to published clinical parameters. The evaluation revealed an important methodological finding, that apparent robustness can mask majority-class defaulting, which has implications beyond this specific task. All experiments are CPU-reproducible with documented software versions and random seeds, and the analysis honestly contextualizes all results against the majority-class baseline.

4.4 Limitations

Several limitations affect the interpretation and generalizability of these findings:

- (1) **Severe class imbalance.** The dataset contains 45 control and 250 AD recordings (85% AD). Raw accuracy is a misleading metric in this setting; a trivial majority-class classifier achieves 84.7%. Macro F1 and balanced accuracy provide more informative evaluation but are limited by the small number of control samples.
- (2) **Single dataset.** All experiments used DementiaBank only. Cross-dataset evaluation (train on DementiaBank, test on the balanced ADReSS dataset¹⁹) is the natural next step but requires a data use agreement not yet secured.
- (3) **Controlled simulations, not domain shifts.** The 8 acoustic transformations simulate specific recording conditions but do not test generalization across patient populations, clinical protocols, or recording equipment. They should not be interpreted as evidence of cross-domain robustness.
- (4) **Small test set.** The 59-sample test set limits statistical precision. With approximately 9 control samples in the test set, per-class metrics for the control class have wide uncertainty. Calibration estimates from small test sets are inherently unstable; the ECE of 0.042 should not be generalized beyond this evaluation.

(5) **No subgroup fairness analysis.** With fewer than 30 samples per demographic subgroup, we could not perform meaningful fairness evaluation across age, gender, or education groups. Prior work documents 15–20 point accuracy disparities across such groups.^{20,23}

(6) **No longitudinal evaluation.** We used cross-sectional data. Whether these classifiers detect disease progression over time is unknown.

(7) **Default hyperparameters.** We retained default settings to avoid overfitting without a validation set, but this means the reported performance may not represent the achievable ceiling with proper tuning on a larger dataset.

(8) **Speaker overlap.** Speaker 002 has one recording in train and one in test. This is a confirmed minor speaker-level leak. Speakers 172 and 299 have both recordings in train only. Because the overlap involves a single AD recording, its potential impact is limited in scope, but it cannot be quantified without a speaker-level re-analysis. Future work should enforce strict speaker-level splitting.

4.5 Future Research

Four directions follow directly from this work. First, evaluation on the balanced ADReSS dataset to assess performance without class imbalance confounding. Second, full MAML¹⁰ implementation for improved few-shot adaptation. Third, evaluation using strict speaker-level splits and larger balanced datasets. Fourth, fairness-aware training following Zhao et al.³⁶ to enforce demographic parity.

5. Conclusion

Using DementiaBank (295 recordings from 292 speakers; 45 control, 250 AD) and 1,589 acoustic features combining hand-crafted descriptors with wav2vec 2.0 and HuBERT embeddings, we established baselines of 81.4% raw accuracy (macro F1: 0.594) with low calibration error (ECE = 0.042) on clean data. The macro F1 confirms meaningful discrimination beyond trivial majority-class prediction (which achieves 84.7% raw accuracy but only 0.459 macro F1). Our controlled acoustic simulation framework revealed that under perturbation, classifiers converged to majority-class accuracy (84.7%), suggesting loss of discriminative ability rather than genuine acoustic robustness. This methodological finding, that apparent stability in raw accuracy can mask degradation in balanced performance, is the study's most important contribution and applies broadly to imbalanced clinical classification tasks. The evaluation protocol, baseline results, and the majority-class defaulting caution provide a foundation for future work on balanced datasets, fairness-aware approaches,³⁶ and meta-learning methods¹⁰ in clinical speech AI.

Declarations

Abbreviations

AD: Alzheimer's disease; ECE: Expected Calibration Error; F0: fundamental frequency; HNR: harmonic-to-noise ratio; MAML: Model-Agnostic Meta-Learning; MFCC: Mel-frequency cepstral coefficient; PCA: principal component analysis; RF: Random Forest; RMS: root mean square; RT60: reverberation time (60 dB decay); SNR: signal-to-noise ratio; XGB: XGBoost.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

The author declares no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Ethics Statement

This study is a secondary computational analysis of an existing, de-identified, controlled-access dataset. DementiaBank is distributed by TalkBank at Carnegie Mellon University under a signed data use agreement. No new participants were recruited, no interaction or intervention with participants occurred, and the investigator did not receive directly identifiable private information. No additional IRB approval was obtained for this secondary analysis.

Data and Code Availability

The DementiaBank Pitt Corpus is available through TalkBank (<https://dementia.talkbank.org>) under a signed data use agreement. The dataset is controlled-access and cannot be redistributed by the author under the terms of the agreement. Full source code, fixed train/test split indices, domain-transformation scripts, and evaluation code are available in a private GitHub repository for editorial review and will be made publicly available upon publication.

AI Use Statement

The author used Claude (Anthropic) for minor formatting assistance. All research, analysis, writing, and verification were performed by the author, who takes full responsibility for the accuracy and integrity of the manuscript.

Author Contributions

A.A. Singh: conceptualization, methodology, software, investigation, formal analysis, data curation, writing (original draft), writing (review and editing), visualization.

Acknowledgements

The author thanks the faculty mentors who provided guidance during this project.

Prior Publication Statement

This manuscript has not been previously published, is not under review at any other venue, and has not been posted as a preprint. No substantially overlapping manuscript exists.

References

- [1] Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatol.* 2018;154(11):1247-1248.
- [2] Alzheimer's Association. 2020 Alzheimer's disease facts and figures. *Alzheimers Dement.* 2020;16(3):391-460.
- [3] Baevski A, Zhou Y, Mohamed A, Auli M. wav2vec 2.0: a framework for self-supervised learning of speech representations. In: *NeurIPS*; 2020.
- [4] Balagopalan A, Eyre B, Rudzicz F, Novikova J. To BERT or not to BERT: comparing speech and language-based approaches for Alzheimer's disease detection. In: *INTERSPEECH*; 2020.
- [5] Becker JT, Boiler F, Lopez OL, Saxton J, McGonigle KL. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Arch Neurol.* 1994;51(6):585-594.
- [6] Boersma P, Weenink D. Praat: doing phonetics by computer. Available at: <http://www.praat.org/>. Accessed 2026.
- [7] Brookmeyer R, Abdalla N, Kawas CH, Corrada MM. Forecasting the prevalence of preclinical and clinical Alzheimer's disease in the United States. *Alzheimers Dement.* 2018;14(2):121-129.
- [8] Cummins N, Baird S, Schuller BW. Speech analysis for health: current state-of-the-art methodologies. *Comput Speech Lang.* 2022;72:101298.
- [9] Eyben F, Scherer KR, Schuller BW, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans Affect Comput.* 2016;7(2):190-202.
- [10] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: *ICML*; 2017.
- [11] Fraser KC, Meltzer JA, Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. *J Alzheimers Dis.* 2016;49(2):407-422.
- [12] Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks. *J Mach Learn Res.* 2016;17(1):2096-2030.
- [13] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: *ICML*; 2017.
- [14] Hsu WN, Bolte B, Tsai YHH, et al. HuBERT: self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans Audio Speech Lang Process.* 2021;29:3451-3460.
- [15] König A, Satt A, Sorin A, et al. Automatic speech analysis for the assessment of patients with pre-dementia and Alzheimer's disease. *Alzheimers Dement Diagn Assess Dis Monit.* 2015;1(1):112-124.
- [16] Kouw WM, Loog M. A review of domain adaptation without target labels. *IEEE Trans Pattern Anal Mach Intell.* 2021;43(3):766-785.

- [17] Livingston G, Huntley J, Sommerlad A, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet*. 2020;396(10248):413-446.
- [18] de la Fuente Garcia S, Ritchie C, Luz S. Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review. *J Alzheimers Dis*. 2020;78(4):1547-1574.
- [19] Luz S, Haider F, de la Fuente S, Fromm D, MacWhinney B. Alzheimer's dementia recognition through spontaneous speech: the ADReSS Challenge. In: *INTERSPEECH*; 2020:2172-2176.
- [20] Manly JJ, Mayeux R. Ethnic differences in dementia and Alzheimer's disease. In: *Critical Perspectives on Racial and Ethnic Differences in Health in Late Life*. Washington: National Academies Press; 2004:95-141.
- [21] McFee B, Raffel C, Liang D, et al. librosa: audio and music signal analysis in Python. In: *Proc Python in Science Conf*; 2015.
- [22] Medela A, Picon A. Few-shot learning in histopathological images: reducing the need of labeled data on biological datasets. In: *IEEE ISBI*; 2019.
- [23] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453.
- [24] Peng X, Bai Q, Xia X, et al. Moment matching for multi-source domain adaptation. In: *ICCV*; 2019.
- [25] Sagawa S, Koh PW, Hashimoto TB, Liang P. Distributionally robust neural networks. In: *ICLR*; 2020.
- [26] Schuller B, Steidl S, Batliner A, et al. The INTERSPEECH 2013 computational paralinguistics challenge. In: *INTERSPEECH*; 2013.
- [27] Searle T, Ibrahim Z, Dobson R. Comparing natural language processing techniques for Alzheimer's dementia prediction in spontaneous speech. In: *INTERSPEECH*; 2020.
- [28] Sensoy M, Kaplan L, Kandemir M. Evidential deep learning to quantify classification uncertainty. In: *NeurIPS*; 2018.
- [29] Seyyed-Kalantari S, Zhang H, McDermott MB, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med*. 2021;27(12):2176-2182.
- [30] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56.
- [31] Tóth L, Hoffmann I, Gosztolya G, et al. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Curr Alzheimer Res*. 2018;15(2):130-138.
- [32] Wang M, Deng W. Deep visual domain adaptation: a survey. *Neurocomputing*. 2018;312:135-153.
- [33] Wang D, Shelhamer E, Liu S, et al. Tent: fully test-time adaptation by entropy minimization. In: *ICLR*; 2021.
- [34] Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: *NeurIPS*; 2014.

- [35] Yuan J, Bian Y, Cai X, et al. Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease. In: INTERSPEECH; 2020.
- [36] Zhao C, Jiang K, Wu X, Wang H, Khan L, Grant C, Chen F. Algorithmic fairness generalization under covariate and dependence shifts simultaneously. In: Proc 30th ACM SIGKDD (KDD'24); 2024:4419-4430.
- [37] Zhou K, Liu Y, Zhou B, Qiao T. Domain generalization: a survey. IEEE Trans Pattern Anal Mach Intell. 2022.