Simple Beats Complex: Benchmarking Self-Supervised Speech Models as Screening Indicators for Depression

Shubham Thakur
Indian Institute of Technology Delhi
New Delhi, India
jtm242078@iitd.ac.in

Abstract

Speech carries potential indicators of depression—flattened intonation, reduced energy, longer pauses, and slower speaking rate—which makes it a candidate signal for automated screening support. Self-supervised speech models such as WavLM, HuBERT, and wav2vec 2.0 can pick up on these patterns directly from raw audio, yet it remains unclear which model is most effective, how much fine-tuning is actually needed, and whether pairing them with traditional handcrafted features offers any benefit. We address these questions through a systematic comparison of 16 configurations on the English-language DAIC-WOZ corpus (189 clinical interviews), evaluated under speaker-independent 5-fold cross-validation. The results paint a clear picture: (1) partially fine-tuning WavLM-Base—unfreezing just the last 2 of its 12 layers—yields the strongest performance at 0.641 F1, outperforming WavLM-Large, HuBERT, and wav2vec 2.0; (2) fusing SSL representations with handcrafted features consistently degrades performance across all three fusion strategies we tested; (3) MUSAN noise augmentation provides no benefit to any model; and (4) the system detects severe depression (71%) more reliably than mild cases (59%), and performs notably better on female speakers (F1: 0.658) than male speakers (F1: 0.629). We additionally explore a hierarchical layer grouping approach drawing on recent insights about multi-layer SSL representations. While this module successfully learns to separate acoustic from semantic layers, it does not improve upon straightforward fine-tuning. A thorough error analysis uncovers that false negatives tend to cluster near the diagnostic boundary and disproportionately involve male speakers—a pattern that raises important fairness questions for any potential clinical use. Taken together, our findings suggest that when working with small clinical datasets, keeping things simple—adapting a compact model with minimal architectural overhead—is the most effective strategy.

1 Introduction

Depression affects over 280 million people worldwide (World Health Organization, 2023), yet it frequently goes undetected because diagnosis relies on subjective clinical judgment (Cummins et al., 2015). Speech offers a practical route toward automated screening support: depressed individuals tend to speak with less pitch variation, lower energy, more pauses, and a slower rate (Low et al., 2020; Cummins et al., 2015). It is important to note that these markers must differ substantially from the normal individual variation in speech characteristics that exists in the healthy population—a point we return to in the Limitations section.

Two approaches dominate the literature. The first extracts handcrafted features—MFCCs, pitch, energy, spectral descriptors—and feeds them to classifiers like SVMs (Eyben et al., 2016, 2013). The second uses self-supervised learning (SSL) models like wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and WavLM (Chen et al., 2022), which learn speech representations from large unlabeled corpora. A natural idea is to combine both, and several papers propose dual-branch architectures (Zhao et al., 2024). But does this actually help?

We set out to answer this alongside several related questions: Which SSL model works best? How much fine-tuning is needed? Does noise augmentation help? Do larger models perform better? Rather than testing one configuration, we run a controlled benchmark: 16 configurations evaluated under the same 5-fold cross-validation protocol on the English-language DAIC-WOZ corpus (Gratch et al., 2014). This includes 4 SSL backbones, 4 fusion strategies, MUSAN augmentation (Snyder et al., 2015), and a hierarchical layer grouping module inspired by HAREN-CTC (Li et al., 2024). We

also conduct a detailed error analysis to understand where and why the model fails, and discuss the ethical implications throughout the paper (see the Ethics Statement).

Our contributions are as follows. First, we present a controlled benchmark showing that partial fine-tuning of WavLM-Base (0.641 F1) beats all alternatives, including WavLM-Large (0.613), HuBERT (0.622), and wav2vec 2.0 (0.604). Second, we provide evidence that fusing SSL representations with handcrafted features consistently degrades performance across all three fusion architectures, even when pitch is included. Third, we offer a gender and severity analysis showing that the model detects depression better in women (71% recall) than men (44%) and catches severe cases (71%) more reliably than mild ones (59%). Fourth, we present a detailed error analysis showing that false negatives cluster near the PHQ-8 diagnostic threshold (mean score 14.1) and disproportionately affect male speakers. Fifth, we describe a hierarchical layer grouping module that learns to separate WavLM layers into acoustic and semantic groups, validating the hypothesis from Li et al. (2024) in a simpler framework.

We emphasize that our system is intended as a screening support tool that would flag potential cases for professional review—not as a standalone diagnostic instrument. We discuss the conditions under which deployment might be appropriate in the Ethics Statement.

2 Related Work

The DAIC-WOZ corpus is the most widely used benchmark for speech-based depression detection. It contains 189 clinical interviews conducted in English with PHQ-8 labels (Gratch et al., 2014). A systematic review found only 5 of 66 studies on this dataset met basic reproducibility standards (Rutowski et al., 2024). Another study showed that some models exploit interviewer biases rather than learning real depression markers (Burdisso et al., 2024). We use participant-only speech and evaluate with both standard split and 5-fold cross-validation to address both concerns.

Traditional handcrafted feature sets like eGeMAPS (Eyben et al., 2016) and ComParE (Eyben et al., 2013) capture pitch, energy, jitter, shimmer, and spectral properties. When combined with SVMs, they achieve F1 scores of 0.50–0.65 on DAIC-WOZ (Valstar et al., 2016; Rejaibi et al., 2022). These approaches offer interpretability but are inherently limited by what the designer chooses to measure.

Among SSL models, wav2vec 2.0 (Baevski et al., 2020) learns by predicting masked speech segments. HuBERT (Hsu et al., 2021) uses offline clustering for pseudo-labels. WavLM (Chen et al., 2022) adds a denoising objective that preserves speaker characteristics—exactly the properties affected by depression. Fine-tuned wav2vec 2.0 achieves 0.79 F1 on DAIC-WOZ (Zhang et al., 2024), while frozen HuBERT-base reaches approximately 0.55 (Gómez-Zaragozá et al., 2025).

Current state-of-the-art methods include speaker disentanglement approaches reaching 0.80 F1 (Salekin et al., 2023), CNN architectures on Mel spectrograms achieving 0.825 (Triohin et al., 2025), and the HAREN-CTC framework (Li et al., 2024), which uses hierarchical multi-layer SSL features with CTC loss, achieving 0.81 F1 under upper-bound evaluation but only 0.576 under cross-validation—a gap that underscores how evaluation protocol affects reported performance.

3 Method

Figure 1 shows the full architecture. We evaluate configurations of increasing complexity, as described below.

3.1 Data Preparation

From each of the 189 DAIC-WOZ sessions, all conducted in English, we extract participant-only speech using transcript speaker labels, discarding all interviewer utterances to avoid the biases documented by Burdisso et al. (2024). We retain utterances between 1 and 15 seconds in length, select the 20 longest per session, and pad or truncate each to 10 seconds (160,000 samples at 16 kHz). This yields approximately 3,780 utterances across all sessions.

3.2 Handcrafted Features

We extract a 130-dimensional frame-level feature vector using `librosa`, comprising 40 MFCCs plus their first- and second-order derivatives (80 Mel banks, 512-point FFT, 25 ms window, 10 ms hop), pitch via the pYIN algorithm (Mauch and Dixon, 2014), RMS energy, zero-crossing rate, and 7-band spectral contrast. Feature sequences are downsampled to 100 frames per utterance.

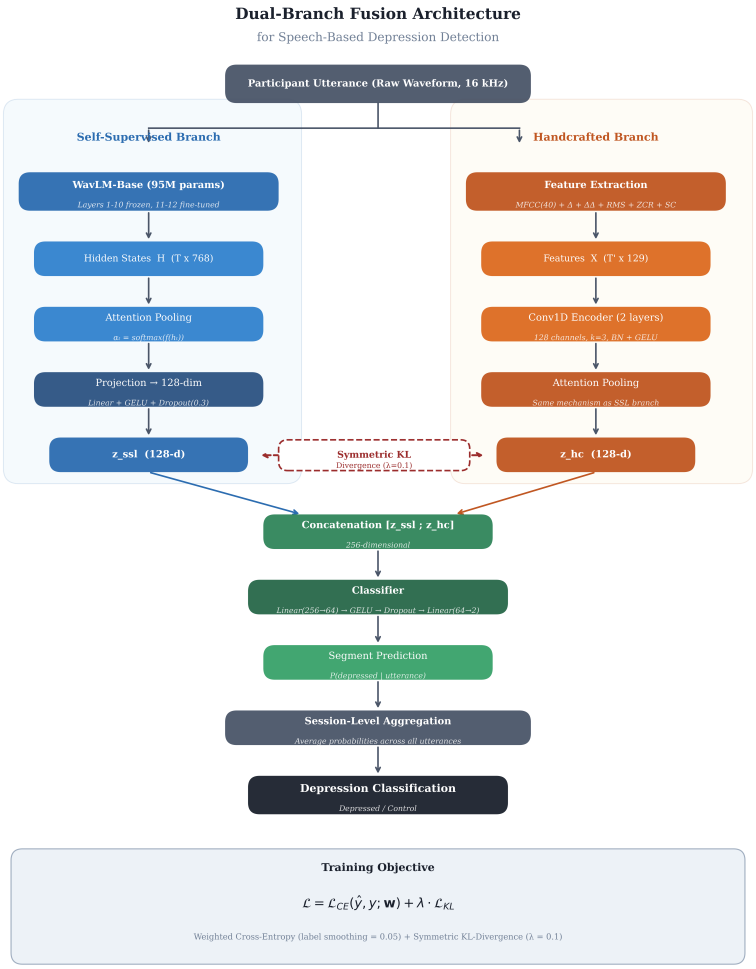

Figure 1: Architecture overview. The SSL branch fine-tunes WavLM (or HuBERT/wav2vec 2.0) with learned attention pooling. The optional handcrafted branch encodes 130-dimensional features (MFCCs + pitch + energy + spectral contrast) via 1D convolutions. Branches can be fused by concatenation, gated fusion, or cross-attention. Our benchmark shows the SSL branch alone gives the best results.

3.3 SSL Models

We evaluate four pre-trained models, each with partial fine-tuning. For all Base models (95M parameters, 12 transformer layers, 768-dim), we freeze all layers except the last two. For WavLM-Large (315M parameters, 24 layers, 1024-dim), we freeze all layers except the last four. We chose to fine-tune only the top layers based on two considerations: first, deeper layers encode more task-relevant semantic and speaker information (Chen et al., 2022; Li et al., 2024), making them better candidates for adaptation; second, freezing most parameters reduces the risk of overfitting on a dataset of only 189 sessions. We did not experiment with unfreezing fewer or more layers, which means our claim about partial fine-tuning being optimal applies specifically to this configuration. Exploring the full spectrum of fine-tuning depths is an important direction for future work.

The output hidden state sequence $\mathbf{H} \in \mathbb{R}^{T \times d}$ is aggregated via learned attention pooling:

$$\alpha_t = \frac{\exp(f(\mathbf{h}_t))}{\sum_{t'} \exp(f(\mathbf{h}_{t'}))}, \qquad \mathbf{z} = \sum_t \alpha_t \mathbf{h}_t \quad (1)$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a two-layer network ($d \to 128 \to 1$, Tanh activation). This mechanism allows the model to weight informative frames more heavily. The pooled representation is classified via Linear(d, 128)–GELU–Dropout(0.5)–Linear(128, 2).

3.4 Fusion Strategies

We combine $\mathbf{z}_{\text{ssl}} \in \mathbb{R}^{128}$ and $\mathbf{z}_{\text{hc}} \in \mathbb{R}^{128}$ using three architecturally distinct methods to en-

sure our conclusions about fusion are not specific to a single approach. In concatenation fusion, the two embeddings are concatenated into a 256-dimensional vector and classified by a two-layer MLP. In gated fusion, a learned gate $\mathbf{g} = \sigma(W[\mathbf{z}_{\text{ssl}}; \mathbf{z}_{\text{hc}}])$ controls the contribution of each branch, producing a fused representation $\mathbf{g} \odot \mathbf{z}_{\text{ssl}} + (1-\mathbf{g}) \odot \mathbf{z}_{\text{hc}}$. In cross-attention fusion, both embeddings form a 2-element sequence processed by 4-head self-attention, allowing each branch to attend to the other. We additionally test symmetric KL-divergence regularization ($\lambda = 0.1$) to encourage the branches to learn complementary rather than redundant features.

3.5 Hierarchical Layer Grouping

Inspired by the HAREN-CTC framework (Li et al., 2024), we propose a simpler alternative that learns to separate WavLM's 13 hidden states into acoustic (shallow) and semantic (deep) groups using a soft assignment matrix $\mathbf{G} \in \mathbb{R}^{13\times 2}$:

$$P_{l,k} = \frac{\exp(G_{l,k})}{\sum_{k'} \exp(G_{l,k'})}, \quad \mathbf{U}^{(k)} = \sum_{l=0}^{12} P_{l,k} \mathbf{H}_l \tag{2}$$

The matrix $\mathbf{G}$ is initialized with exponential decay (biasing shallow layers toward the acoustic group and deep layers toward the semantic group) but is fully learnable during training. The semantic group then queries the acoustic group via cross-attention:

$$\mathbf{F} = \text{softmax}\left(\frac{W_Q \mathbf{U}^{\text{sem}} (W_K \mathbf{U}^{\text{aco}})^\top}{\sqrt{d}}\right) W_V \mathbf{U}^{\text{aco}} \tag{3}$$

followed by a residual connection and layer normalization. Unlike HAREN-CTC, our approach does not rely on CTC loss or HuBERT pseudo-labels, making it substantially simpler to implement and train.

3.6 Training Details

All models are optimized with AdamW (Loshchilov and Hutter, 2019) using differential learning rates: 5×10^{-6} for SSL parameters and 10^{-3} for the classification head. We train for 15 epochs with batch size 8, FP16 mixed precision, and gradient clipping at norm 1.0. Class weights (0.695 for control, 1.783 for depressed) address the imbalanced label distribution, and label smoothing of 0.05 prevents overconfident predictions. At inference time, session-level predictions are obtained by averaging the softmax probabilities across all utterances from a given session. When MUSAN augmentation (Snyder et al., 2015) is applied, each training utterance has a 50% chance of being corrupted with a randomly selected noise, music, or speech sample at a random SNR between 5 and 20 dB.

	Ctrl	**Dep**	**Total**	**Dep Rate**
Male	77	25	102	24.5%
Female	56	31	87	35.6%
Severity (depressed only):				
Mild (10–14)	–	29		51.8%
Moderate (15–19)	–	20		35.7%
Severe (20+)	–	7		12.5%

Table 1: DAIC-WOZ demographics and severity.

4 Experimental Setup

We use the DAIC-WOZ corpus (Gratch et al., 2014), which contains 189 clinical interviews conducted in English at the University of Southern California. Participants were assessed using the PHQ-8 questionnaire, with a score of 10 or above indicating depression. This yields 56 depressed and 133 control participants. We chose DAIC-WOZ because it is the most widely used benchmark in this area, enabling direct comparison with prior work. We acknowledge that evaluating on a single dataset limits the generalizability of our conclusions; cross-corpus evaluation on datasets such as MODMA (Cai et al., 2022) (Mandarin) or E-DAIC (English) would strengthen our claims and is planned as future work. Table 1 summarizes the demographic characteristics.

We evaluate under two protocols. First, the standard split (train: 107, dev: 35, test: 47 sessions), with model selection based on dev F1, enables comparison with published results. Second, 5-fold speaker-independent cross-validation over all 189 sessions provides a more robust estimate with error bars. We report session-level macro F1 throughout. All experiments were conducted on a single NVIDIA RTX PRO 6000 GPU (96 GB) with fixed random seed 42.

5 Results

5.1 SSL Model Comparison

Table 2 compares the four SSL models. WavLM-Base achieves the highest mean F1 (0.641), likely because its denoising pre-training objective preserves speaker-level characteristics relevant to de-

Model	Params	F1	Std
wav2vec 2.0-Base	95M	.604	.099
HuBERT-Base	95M	.622	.052
WavLM-Large	315M	.613	.137
WavLM-Base	**95M**	**.641**	**.067**

Table 2: SSL model comparison (5-fold CV).

Model	No MUSAN	+MUSAN	Δ
wav2vec 2.0	.604	.554	−5.0pp
HuBERT	.622	.622	0.0pp
WavLM-Base	.641	.635	−0.6pp

Table 3: Effect of MUSAN augmentation.

Method	F1	Std
SSL-only (no fusion)	**.641**	**.067**
+ HC, Concatenation	.545	.065
+ HC, Concat + KL reg.	.544	.044
+ HC, Gated fusion	.560	.070
+ HC, Cross-attention	.583	.074

Table 4: Fusion methods (WavLM-Base FT; HC features include pitch).

pression. WavLM-Large performs worse (0.613) with 3.5× higher variance across folds, suggesting that the larger model overfits when training data is limited to approximately 150 sessions.

5.2 MUSAN Augmentation

MUSAN noise augmentation does not improve any model (Table 3). It actually degrades wav2vec 2.0 by 5 percentage points. Depression-related speech markers are subtle—reduced energy, slight prosodic flattening—and adding noise at 5–20 dB can mask these signals. Moreover, SSL models were already pre-trained on diverse audio conditions, so additional noise robustness training provides little benefit for this task.

5.3 Fusion with Handcrafted Features

Every fusion method tested performs worse than SSL alone (Table 4 and Figure 2). The best fusion approach (cross-attention, 0.583) still falls 5.8 percentage points below the SSL-only baseline. KL-divergence regularization has no measurable effect (−0.1pp). We interpret this as evidence that WavLM already captures the information present in MFCCs and pitch from raw audio. Adding the handcrafted branch introduces redundant features and extra trainable parameters that increase overfitting risk on this small dataset.

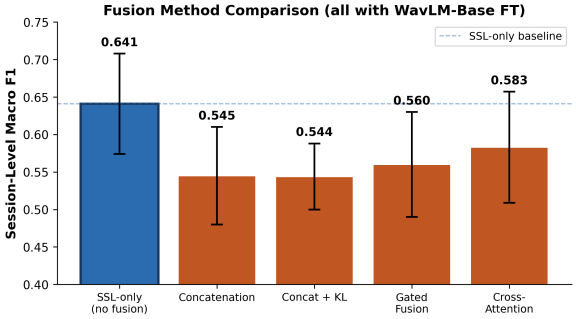

Figure 2: All fusion methods (orange) fall below the SSL-only baseline (blue). Dashed line marks 0.641.

Method	F1	Prec.	Dep. Rec.
HC-only (SVM)	.496	—	—
Frozen WavLM (SVM)	.482	—	—
WavLM-Base FT	**.659**	**.661**	**.714**

Table 5: Standard DAIC-WOZ test split (47 sessions).

5.4 Complete Benchmark

Figure 3 ranks all 16 configurations. Fine-tuned SSL-only models cluster at the top, while fusion approaches cluster at the bottom. We note that several error bars overlap, particularly among mid-ranking methods. The statistical significance of key comparisons is assessed formally in Section 5.8.

5.5 Standard Split

On the standard DAIC-WOZ test split (47 sessions), our model achieves 0.659 F1 with 71.4% depressed recall (Table 5), correctly identifying 10 of 14 depressed participants.

5.6 Comparison with Published Work

Our cross-validation result (0.641) exceeds the CV result reported for HAREN-CTC (0.576; Li et al., 2024) despite using a considerably simpler architecture (Table 6). Standard-split results are not directly comparable to upper-bound evaluations that select the best epoch on the development set.

5.7 Hierarchical Layer Grouping

Our hierarchical model achieves 0.600 F1—above all fusion methods but below simple fine-tuning (0.641). While it does not improve the overall score, the learned layer assignments (Figure 4) reveal meaningful structure: layers 0–1 are strongly assigned to the acoustic group (weights 0.78, 0.70), layers 2–4 show mixed assignments (∼0.53), and layers 5–12 increasingly map to the semantic group (0.46→0.23 acoustic weight). This confirms, with-

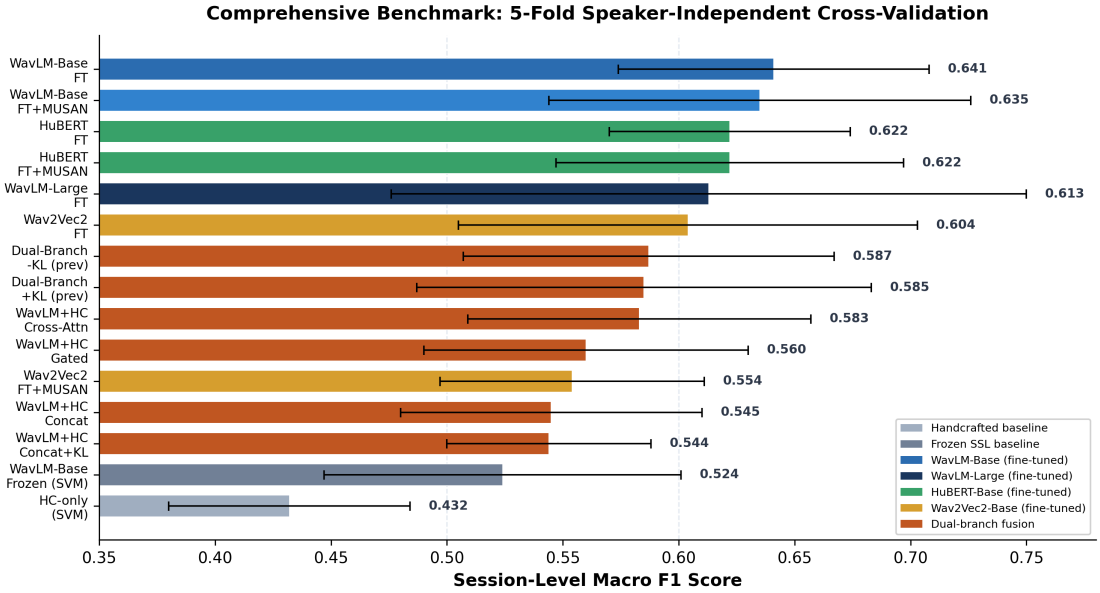

Figure 3: All 16 configurations ranked by 5-fold CV session-level F1. Error bars show standard deviation. While some differences between adjacent methods are not statistically significant (see Table 7), the separation between SSL-only models (top) and fusion methods (bottom) is consistent.

Method	F1	Eval
AVEC baseline (Valstar et al., 2016)	.530	Std
MFCC+RNN (Rejaibi et al., 2022)	.649	Std
wav2vec 2.0+CNN (Zhang et al., 2024)	.790	Std
Speaker disent. (Salekin et al., 2023)	.800	Std
HAREN-CTC (Li et al., 2024)	.810	UB
HAREN-CTC (Li et al., 2024)	.576	CV
Ours (WavLM-Base FT)	.659	Std
Ours (WavLM-Base FT)	**.641**	**CV**

Table 6: Published audio-only DAIC-WOZ results. UB = upper-bound. CV = cross-validation. Std = standard split.

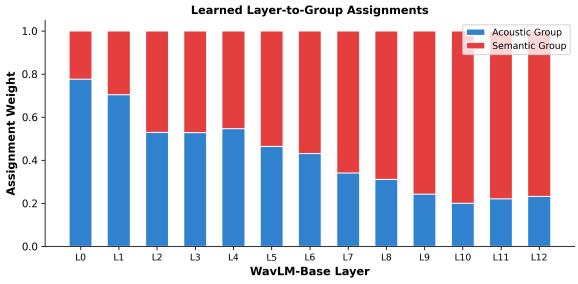

Figure 4: Learned layer-to-group assignments. The model discovers the expected acoustic→semantic gradient without explicit supervision.

out explicit supervision, that WavLM's shallow layers encode prosodic patterns while deeper layers encode speaker-level and semantic information—consistent with the findings of Li et al. (2024) and Chen et al. (2022).

WavLM-Base FT vs.	t	p
HC-only (SVM)	+9.47	.0007*
Frozen WavLM (SVM)	+4.76	.009*
WavLM+HC concat	+3.69	.021*
WavLM+HC cross-attn	+1.74	.157
WavLM-Large	+0.07	.946
Hierarchical (ours)	+0.54	.616

Table 7: Paired t-tests across 5 folds. * = $p < 0.05$.

5.8 Statistical Significance

Table 7 presents paired t-tests across the 5 CV folds. Fine-tuning significantly outperforms frozen representations ($p = .009$) and concatenation fusion ($p = .021$). The differences between WavLM-Base FT and cross-attention fusion, WavLM-Large, and our hierarchical model are not statistically significant at $\alpha = 0.05$, though the mean differences are consistent in direction across all folds. Given the small number of folds ($n = 5$), statistical power is limited.

5.9 Gender and Severity Analysis

The model catches depression in women substantially better than in men (recall: 71% vs. 44%; Table 8). This disparity likely stems from two factors: the training data contains proportionally more depressed women (35.6% vs. 24.5%), and prior clinical research suggests that women may exhibit more pronounced vocal changes when depressed (Cummins et al., 2015). The model also catches severe

	Sessions	F1	Rec.	Prec.
All	189 (56 dep)	.657	.589	.493
Male	102 (25 dep)	.629	.440	.440
Female	87 (31 dep)	.658	.710	.524

Table 8: Gender-stratified performance (all 5 folds).

Severity	PHQ-8	Detected	Rate
Mild	10–14	17/29	59%
Moderate	15–19	11/20	55%
Severe	20+	5/7	71%

Table 9: Detection rate by depression severity.

Category	Count	Rate
By proximity to threshold (depressed only)		
Near-threshold (PHQ-8: 10–12)	13/24	54%
Above threshold (PHQ-8: 13+)	20/32	63%
False negative gender breakdown		
Male false negatives	14/23	61%
Female false negatives	9/23	39%
Model confidence		
Mean prob. (true positives)	0.664	
Mean prob. (false negatives)	0.283	

Table 10: Error analysis across 5-fold CV (189 sessions).

depression (71%) more reliably than mild cases (59%; Table 9). This is expected: stronger symptoms produce more detectable speech changes, while mild depression (PHQ-8 scores of 10–12) may produce vocal patterns that are indistinguishable from normal individual variation in healthy speakers. The low recall for male speakers raises concerns about equitable performance that must be addressed before any clinical use (see the Ethics Statement).

5.10 Error Analysis

To understand where the model fails, we examine the 23 false negatives and 8 false positives across all 5 CV folds. Table 10 summarizes the patterns.

False negatives cluster near the diagnostic threshold: the mean PHQ-8 score of missed depressed sessions is 14.1 (median: 13.0). Sessions with PHQ-8 scores of 10–12 are detected at only 54%, while those above 13 are detected at 63%. This is not surprising: mild depression produces subtler speech changes, and the binary threshold at PHQ-8=10 groups together individuals whose vocal markers may be indistinguishable from healthy controls.

Male false negatives are disproportionately common: of the 23 missed depressed sessions, 14 (61%) are male speakers. Several male participants with moderate-to-severe depression (PHQ-8 $\geq$ 15) were missed with very low predicted probabilities ($<$0.2), suggesting that the model has not learned male-specific depression markers effectively.

The confidence gap between true positives (mean probability 0.664) and false negatives (0.283) suggests that a clinical deployment could productively flag low-confidence predictions for human review rather than relying on a fixed decision threshold. We note that low precision is also a concern: our model produces 8 false positives across the 189 sessions. While the clinical cost of a false positive (unnecessary follow-up assessment) is generally lower than that of a false negative (missed depression), it still has implications for patient anxiety and healthcare resource allocation.

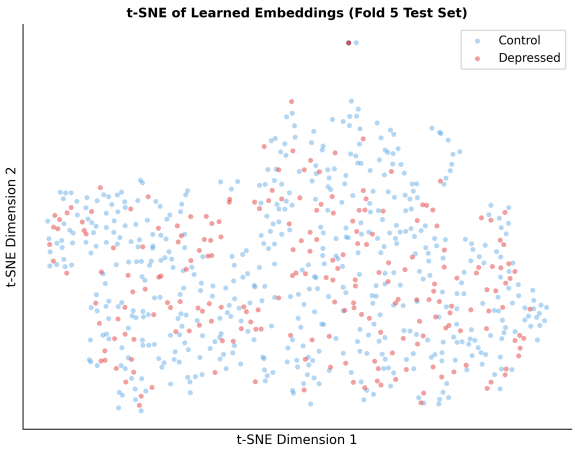

Figure 5: t-SNE of learned embeddings. Partial but imperfect separation between depressed (red) and control (blue) utterances.

5.11 Embedding Visualization

Figure 5 shows a t-SNE visualization of utterance-level embeddings from the hierarchical model (fold 5 test set). Depressed utterances form a loose cluster offset from controls, with substantial overlap. This is expected: depression cues are subtle, vary across individuals, and are not present in every utterance of a session.

6 Discussion

Our results yield several insights. Frozen WavLM captures general speech structure but not the subtle paralinguistic differences that characterize depressed speech. Updating just the top 2 layers (20% of all parameters) is sufficient to adapt the

model, producing an 11.7 percentage point gain over frozen representations ($p = .009$). This is consistent with findings from other clinical speech tasks where modest fine-tuning outperforms both frozen representations and full fine-tuning on small datasets.

The consistent failure of all three fusion architectures is perhaps our most noteworthy finding. WavLM processes raw waveforms and implicitly learns representations that capture the same acoustic properties measured by MFCCs and spectral descriptors. The handcrafted branch therefore adds redundant information and extra trainable parameters that increase overfitting risk on a dataset of only 150 training sessions. The fact that three architecturally distinct fusion methods all show this pattern makes it unlikely that the problem is specific to any one fusion approach.

WavLM-Large has $3.3\times$ more parameters than WavLM-Base but is trained on the same amount of downstream data. Its $3.5\times$ higher cross-fold variance confirms overfitting. This practical finding suggests that researchers working with small clinical datasets should prefer compact models.

The hierarchical layer grouping module correctly discovers the acoustic-semantic hierarchy of WavLM's layers (Figure 4), providing independent validation of the core hypothesis underlying HAREN-CTC (Li et al., 2024). However, it does not beat simple fine-tuning, likely because the cross-attention mechanism adds trainable parameters without sufficient data to learn from. Combining this grouping with CTC-based temporal regularization, as in HAREN-CTC, may help and is a direction for future work.

The error analysis reveals two systemic failure modes with implications for deployment. First, the binary PHQ-8=10 threshold creates a boundary where depressed and healthy speech may be genuinely indistinguishable, suggesting that a regression or ordinal approach might better capture the severity spectrum. Second, the 27-percentage-point gender gap in recall is not simply a matter of training data imbalance; even male participants with PHQ-8 scores above 15 are sometimes missed with high confidence, suggesting that the acoustic correlates of depression differ across genders in ways the model does not capture.

7 Conclusion

We benchmarked 16 configurations for speech-based depression screening support on DAIC-WOZ under speaker-independent 5-fold cross-validation. The central finding is practical: partially fine-tuning WavLM-Base (0.641 F1) outperforms every alternative we tested, including larger models, handcrafted feature fusion, noise augmentation, and hierarchical architectures.

Three results carry broader implications. First, the failure of all three fusion architectures suggests that when an SSL model has been pre-trained on sufficient data, handcrafted features add redundancy rather than complementary information. Second, the 27-percentage-point gender gap in depression recall is not merely a limitation but a barrier to equitable deployment. Third, our error analysis reveals that missed cases cluster near the PHQ-8 diagnostic threshold, suggesting that binary classification may be fundamentally mismatched with the continuous nature of depression severity. For researchers working with small clinical datasets, our results offer concrete guidance: use a compact SSL model, fine-tune only the top layers, and invest effort in evaluation rigor rather than architectural complexity.

Limitations

All experiments use a single English-language dataset (DAIC-WOZ, 189 sessions). Depression-related speech markers may vary across languages, recording conditions, and clinical contexts. Cross-corpus evaluation on datasets such as MODMA (Cai et al., 2022) or E-DAIC is needed before drawing broader conclusions. Our 130-dimensional handcrafted feature set omits jitter, shimmer, harmonics-to-noise ratio, and formant frequencies—features that prior work has linked to depression (Cummins et al., 2015). Including these could change our conclusion about fusion utility. We tested only one fine-tuning configuration (last 2 or 4 layers); other depths may yield different results. We treat depression as a binary classification (PHQ-8 $\geq$ 10), ignoring severity gradations, and process utterances independently without modeling temporal dynamics across the interview.

Ethics Statement

This system is intended as a screening support tool—not as an autonomous diagnostic instrument. We envision its use in a scenario where

low-confidence predictions are flagged for review by a trained clinician, rather than used to make clinical decisions directly. Based on our results, we would not consider deployment appropriate until the system achieves at minimum: (a) balanced recall across genders, (b) validation on at least one additional dataset in a different language or clinical setting, and (c) demonstrated value in a prospective clinical study with appropriate oversight.

False negatives risk missed diagnoses, while false positives may cause unnecessary anxiety or lead to wasteful follow-up assessments. Both types of error carry costs, and a deployment scenario must carefully calibrate the decision threshold based on the specific clinical context. The DAIC-WOZ dataset was collected under IRB approval with informed consent. Speech data contains biometric information requiring careful handling under regulations such as HIPAA and GDPR. The technology could potentially be misused for non-consensual speaker profiling, workplace surveillance, or insurance discrimination.

Reproducibility

All training code, configuration files, and evaluation scripts are available at `https://github.com/shubham-899844/depression-detection-benchmark`. We report results under both standard split and 5-fold CV, specify all hyperparameters in Section 3.6, use fixed random seeds, and report standard deviations alongside mean scores.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Sergio Burdisso, Matthieu Labeau, Brice Loison, and Juan Zuluaga-Gomez. 2024. On the validity of using the therapist's prompts in automatic depression detection. In *Proceedings of the 6th Clinical Natural Language Processing Workshop (NAACL)*, pages 82–90.

Hanshu Cai, Zhenqin Yuan, Yiwen Gao, Shuting Sun, Na Li, Fuze Tian, Han Xiao, Jianxiu Li, Zhengwu Yang, Xiaowei Li, Qinglin Zhao, Zhenyu Liu, Zhijun Yao, Minqiang Yang, Hong Peng, Jing Zhu, Xiaowei Zhang, Guoping Gao, Fang Zheng, and 8 others. 2022. A multi-modal open dataset for mental-disorder analysis. *Scientific Data*, 9(1):178.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F. Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 835–838.

Patricia Gómez-Zaragozá, Javier Marín-Morales, and Mariano Alcañiz. 2025. Speech and text foundation models for depression detection. In *Proceedings of Interspeech*.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The Distress Analysis Interview Corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 3123–3128.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Yujie Li, Eng Siong Chng, and Cuntai Guan. 2024. Hierarchical self-supervised representation learning for depression detection from speech. *arXiv preprint arXiv:2510.08593*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Daniel M. Low, Kate H. Bentley, and Satrajit S. Ghosh. 2020. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5(1):96–116.

Matthias Mauch and Simon Dixon. 2014. pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 659–663.

Emna Rejaibi, Ali Komaty, Fabrice Meriaudeau, Said Agrebi, and Alice Othmani. 2022. MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*, 71:103107.

Adam Rutowski, Elizabeth Shriberg, and Amir Harati. 2024. A systematic review of depression detection from speech using DAIC-WOZ. *Sensors*, 24.

Asif Salekin, Jeremy W. Eberle, Jeffrey J. Glenn, Bethany A. Teachman, and John A. Stankovic. 2023. Speaker disentanglement for speech-based depression detection. *Computer Speech and Language*, 82:101533.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. MUSAN: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

Vlad Triohin, Mihaela Dinsoreanu, and Camelia Lemnaru. 2025. A CNN-to-SNN approach for depression detection using DAIC-WOZ. *Applied Sciences*, 15.

Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 3–10.

World Health Organization. 2023. Depressive disorder (depression). Fact Sheet.

Xin Zhang, Jintao Zhang, Weiran Xu, and Jun Guo. 2024. Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments. *Scientific Reports*, 14:9596.

Zhiwei Zhao, Yuqing Zhang, and Wei Chen. 2024. Multi-modal approaches for depression detection. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*.