Speaker-Aware Temporal Aggregation Strategies on Segment Representations for Depression Detection in Dyadic Interaction: A Benchmark Study

Anisha Pattanayak[1], Huang-Cheng Chou[2], Shrikanth Narayanan[2], Sudarsana Reddy Kadiri[2]
[1]*Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, USA*
[2]*Signal Analysis and Interpretation Laboratory (SAIL), University of Southern California, USA*

***Abstract*—Speech-based depression detection compresses features from short audio segments into one speaker-level decision, a step called temporal aggregation rarely studied on its own. Most benchmarks fix a single self-supervised encoder and a single hand-picked layer, so a reported gain may reflect the pipeline rather than the aggregation method itself. We introduce DEPOOL, a controlled benchmark that compares six aggregation architectures with six frozen speech backbones on an English and a Mandarin depression corpus, where each configuration learns which backbone layers matter rather than fixing one by hand. Across the resulting 72-configuration grid, a third of configurations collapse into predicting a single class for every speaker, a failure tied to the backbone as much as to the method, and the architecture that is most stable in a single-seed run becomes unreliable when training repeats across seeds. Robustness to backbone and seed, rather than average accuracy across a single pipeline, should be a first-class benchmarking criterion for temporal aggregation in clinical speech.**

***Index Terms*—depression detection, temporal aggregation, pooling, clinical speech, self-supervised learning, speaker-level evaluation, benchmark robustness**

I. Introduction

Automatic depression detection from speech is attracting growing interest as a complement to clinical screening. The World Health Organization estimates that over 280 million people live with depression, fewer than half of whom receive treatment [1], and speech can be collected passively during a routine clinical interaction [3].

Most systems in this space share a common structure where a feature extractor processes fixed-length audio segments, and a classifier turns those segment features into a single speaker-level prediction covering the whole interview. This last step, *aggregation*, is where many of the open design questions in the literature live. A clinical interview is not a bag of interchangeable snippets because psychomotor retardation, a reliable acoustic marker of depression, tends to build across time rather than sit uniformly across a recording [2]. An aggregation method that tracks this structure should, in principle, beat one that averages everything together.

A subtler problem sits underneath this one. Almost every existing comparison of aggregation methods is run on a single Self-Supervised Learning (SSL) backbone, which is a large speech model pre-trained on unlabeled audio, at a single, manually chosen transformer layer. This layer is typically the one the authors' own probing happened to favor. That makes it hard to tell whether an observed advantage belongs to the aggregation architecture itself or to the particular backbone and layer it was paired with. This paper asks whether the relative ranking of temporal aggregation strategies for depression detection holds up once the SSL backbone, the transformer layer, and the corpus are all allowed to vary, or if it is an artifact of one convenient pipeline.

To answer this, we built **DEPOOL** and trained all six aggregation architectures on all six SSL backbones on both E-DAIC [5] and MODMA [6], forming a 72-cell grid. To remove hand-picked layer selection as a confound, every configuration learns a softmax-weighted combination of all hidden layers in its backbone rather than using a single fixed layer. All splits are speaker-independent and use stratified, participant-grouped folds so that no segment from a test-set speaker ever appears in training. The overall protocol follows the layer-featurization and cross-backbone evaluation conventions established by the SUPERB benchmark family [12], [13] so that results reported here can be read alongside and combined with results reported under those same conventions for other paralinguistic tasks. This paper provides several contributions.

- We present a controlled cross-product benchmark of aggregation architecture, SSL backbone, and corpus rather than a single-pipeline comparison.
- We introduce a semi-fine-tuned layer-aggregation protocol that removes hand-picked-layer selection as a confound.
- We report the finding that roughly a third of the 72 configurations collapse into single-class prediction, a phenomenon invisible to any study reporting one backbone, and we show that this collapse is seed-sensitive as well as backbone-dependent.
- We provide strictly speaker-independent, stratified 60/20/20 splits and an open release of the pipeline, splits, and results[1].

II. Related Work

A. Depression detection from speech

Early systems relied on hand-crafted acoustic features, including Mel-frequency cepstral coefficients (MFCCs), pitch,

[1]https://anonymous.4open.science/r/Temporal_Pooling_Benchmark-5F64

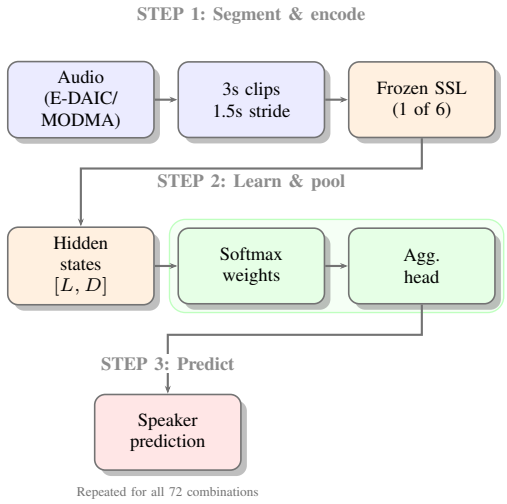

Fig. 1. DEPOOL pipeline: Components are color-coded by training status.

jitter, and shimmer, paired with support vector machines or logistic regression [2], [15]; bidirectional long short-term memory (LSTM) networks later showed that modeling temporal dynamics improves over static summaries [16]. Self-supervised learning (SSL) models such as HuBERT (Hidden-unit Bidirectional Encoder Representations from Transformers, or Hidden-unit BERT) [7], WavLM [4], data2vec [9], and XLS-R (a large-scale cross-lingual speech representation model) [10] learn rich representations from unlabeled audio and consistently outperform hand-crafted pipelines on paralinguistic tasks, including depression detection [17], [18]. The SUPERB (Speech processing Universal PERformance Benchmark) [12] showed that an SSL backbone's apparent quality depends heavily on the task and probing protocol used to evaluate it, and that no single backbone dominates across the board; EMO-SUPERB made a similar reproducibility argument for speech emotion recognition, releasing a unified, leakage-safe benchmark codebase across backbones [13]. Layer-wise studies further show intermediate transformer layers often encode prosodic and affective information better than layers near the input or output [11], [14], motivating our choice to learn layer weights rather than fix them by hand.

B. Segment aggregation in clinical speech

The aggregation step itself has received comparatively little attention, usually confined to a single backbone. Common approaches include majority voting over segment predictions, mean pooling of segment embeddings, or a single recurrent model over the segment sequence [19]. Ringeval et al. [20] fused frame-level predictions with simple statistics; Gaus et al. [21] explored speaker normalization before aggregation without comparing architectures directly; Ma et al. [22] proposed DepAudioNet, a convolutional neural network (CNN) followed by an LSTM, with mean pooling as its aggregation step. Attention-based aggregation has been studied in related paralinguistic tasks, including local attention over LSTM outputs for emotion recognition [23], which relates to the self-attention and GRU-with-attention heads studied here, and NetVLAD (Vector of Locally Aggregated Descriptors) [24] and attentive statistics pooling [25] are standard for speaker embeddings, though neither has previously been evaluated for depression screening across multiple backbones. Transformer pooling with a classification (CLS) token, borrowed from Bidirectional Encoder Representations from Transformers (BERT)-style text classification [26], is likewise beginning to appear in audio tasks without cross-backbone evaluation.

III. The DEPOOL Benchmark

A. Datasets

We use two depression corpora that differ in language, elicitation protocol, and population, so that a finding holding across both corpora is less likely to be an artifact of a single dataset, recording setup, or annotation style.

E-DAIC [5] contains semi-structured English clinical interviews between participants and a virtual agent, labeled with the 8-item Patient Health Questionnaire (PHQ-8), a self-report depression screening scale binarized at PHQ-8 $\geq$ 10 (depressed). After speaker-independent splitting (Section III-F), our partition contributes 122 unique participants, divided into a training subset (75 participants) used to fit model parameters, a validation subset (24 participants) used only for checkpoint selection, and a held-out test subset (23 participants: 17 healthy, 6 depressed) used exactly once for final evaluation and never seen during model or hyperparameter selection. This is a participant subset with complete audio and labels after preprocessing, re-split rather than using the official challenge partition, so absolute numbers are not directly comparable to challenge leaderboards.

MODMA [6] is a Mandarin corpus of clinically diagnosed major depressive disorder (MDD) patients and matched healthy controls (HC), recorded during structured interview and reading tasks and diagnosed by clinicians rather than by self-report. After the same splitting procedure, it contributes 52 participants: a training subset of 31, a validation subset of 11, and a held-out test subset of 10 (5 healthy, 5 depressed), with the same three-way training/validation/test roles as E-DAIC.

B. Segmentation and Preprocessing

Participant-only speech is isolated from each interview (interviewer or virtual-agent turns are excluded, since only the participant's own speech is diagnostically relevant) and resampled to 16 kHz mono to standardize sampling rate across the two corpora's original recording formats. We slide a 3-second window across each participant's turn with a 1.5-second stride, i.e., 50% overlap between consecutive windows; 3 seconds was chosen as long enough to contain several words of connected speech and short enough that the self-supervised backbone's fixed-length input still corresponds to a single, roughly stationary prosodic unit, while the 50% overlap smooths artifacts at arbitrary window boundaries. Turns shorter than one window are dropped, since they cannot fill even one clip. Each clip is zero-mean, unit-variance normalized, to remove per-recording loudness and channel-level differences between the two corpora's collection setups, and cropped or zero-padded to exactly 48,000 samples (3 s at 16 kHz) before being passed to the backbone. An interview

is thus represented downstream as an ordered sequence of 3-second clips, which preserves the coarse temporal ordering of the session while keeping each clip short enough to be treated as one token by the sequence models of Section III-E.

C. SSL Backbones

DEPOOL evaluates six frozen self-supervised learning (SSL) speech backbones (Appendix Table IV): two size variants of WavLM, HuBERT (Hidden-unit BERT), Wav2Vec2-Robust (a noise- and channel-robust variant of wav2vec 2.0, itself a contrastive self-supervised speech encoder), Data2Vec-Audio (an encoder trained with the general-purpose data2vec self-distillation objective), and XLS-R (a large, multilingual cross-lingual speech representation model). All six are large Transformer networks pre-trained on thousands of hours of unlabeled speech audio. "Frozen" means their weights are never updated for our task; only the aggregation head (Section III-E) and the layer-aggregation weights (Section III-D) are trained on top of them. Two checkpoints (HuBERT-Large, Data2Vec-Audio-Large) are automatic-speech-recognition (ASR)-fine-tuned rather than purely self-supervised, meaning they were further trained to transcribe speech to text after self-supervised pretraining; since this fine-tuning reshapes upper layers toward phonetic content, part of the backbone effect we observe may reflect this fine-tuning distinction rather than the pretraining objective alone, a confound our learned layer weighting mitigates but does not remove.

For every clip, we extract the full stack of hidden states and mean-pool each layer over time, giving a $[L, D]$ representation per clip, where L is the number of hidden-state layers (Appendix Table IV) and D is the backbone's hidden size.

D. Semi-Fine-Tuned Layer Aggregation

Different backbone layers capture different information, and picking one by hand biases the comparison. Instead, each configuration learns a soft, weighted average of all L layers (following the featurizer used in SUPERB-style benchmarks [12], which keeps our protocol directly comparable to that established benchmarking convention). For clip t, with per-layer vectors $\{\mathbf{h}_t^{(1)}, \ldots, \mathbf{h}_t^{(L)}\}$, $\mathbf{h}_t^{(\ell)} \in \mathbb{R}^D$:

$$\tilde{\mathbf{h}}_t = \sum_{\ell=1}^{L} \text{softmax}(w_\ell)\, \mathbf{h}_t^{(\ell)}, \qquad \mathbf{z}_t = \mathbf{W}_p \tilde{\mathbf{h}}_t + \mathbf{b}_p \quad (1)$$

, where the symbols are defined as follows:

- $t \in \{1, \ldots, T\}$ indexes the 3-second clip within an interview;
- $\ell \in \{1, \ldots, L\}$ indexes the backbone's hidden-state layer, with L given per backbone in Appendix Table IV;
- $\mathbf{h}_t^{(\ell)} \in \mathbb{R}^D$ is the time-mean-pooled hidden state of layer ℓ for clip t, with hidden size D given per backbone in Appendix Table IV;
- $w_\ell \in \mathbb{R}$ is one learnable scalar weight per layer, shared across all clips and all interviews for a given backbone;
- $\text{softmax}(w_\ell) = \exp(w_\ell) / \sum_{\ell'=1}^{L} \exp(w_{\ell'})$ normalizes the L layer weights to be non-negative and sum to one;
- $\tilde{\mathbf{h}}_t \in \mathbb{R}^D$ is the resulting layer-weighted combination of hidden states for clip t;
- $\mathbf{W}_p \in \mathbb{R}^{256 \times D}$ is a learnable linear projection matrix, shared across all six backbones' downstream heads;
- $\mathbf{b}_p \in \mathbb{R}^{256}$ is the corresponding learnable bias term;
- $\mathbf{z}_t \in \mathbb{R}^{256}$ is the resulting clip embedding passed to the aggregation architectures of Section III-E.

Only $\{w_\ell\}$ (learned separately per backbone) and $\mathbf{W}_p, \mathbf{b}_p$ (shared across backbones) are trained here. This is what makes the benchmark controlled: every aggregation architecture downstream of Eq. (1) sees a 256-dimensional clip embedding regardless of the backbone, so differences reflect what the frozen representations encode, not incidental differences in dimensionality.

E. Aggregation Architectures

Each head takes the ordered clip embeddings $\{\mathbf{z}_1, \ldots, \mathbf{z}_T\}$, $\mathbf{z}_t \in \mathbb{R}^{256}$, of one interview ($T$ the padded clip count) and a binary validity mask $\mathbf{m} \in \{0, 1\}^T$ (where $m_t = 1$ marks a real, non-padded clip and $m_t = 0$ marks padding), and compresses them into a single speaker vector that a classifier maps to depressed vs. healthy. *Mean Pooling.* $\mathbf{p}_{\text{mean}} = \frac{1}{|\mathcal{S}|} \sum_{t \in \mathcal{S}} \mathbf{z}_t$, where $\mathcal{S} = \{t : m_t = 1\}$ is the set of valid (non-padded) clip indices and $|\mathcal{S}|$ its cardinality. *Statistical Pooling.* Concatenates four summary statistics computed over valid clips $\{\mathbf{z}_t\}_{t \in \mathcal{S}}$: the element-wise mean $\mu \in \mathbb{R}^{256}$, standard deviation $\sigma \in \mathbb{R}^{256}$, maximum $\max \in \mathbb{R}^{256}$, and minimum $\min \in \mathbb{R}^{256}$, giving $\mathbf{p}_{\text{stat}} = [\mu; \sigma; \max; \min] \in \mathbb{R}^{1024}$. *Self-Attention Pooling.* Learns a per-clip weight α_t:

$$\alpha_t = \frac{\exp(\mathbf{v}^\top \tanh(\mathbf{W}\mathbf{z}_t + \mathbf{b}))m_t}{\sum_{t'} \exp(\mathbf{v}^\top \tanh(\mathbf{W}\mathbf{z}_{t'} + \mathbf{b}))m_{t'}}, \ \mathbf{p}_{\text{attn}} = \sum_{t \in \mathcal{S}} \alpha_t \mathbf{z}_t \quad (2)$$

, where $\mathbf{W} \in \mathbb{R}^{128 \times 256}$ and $\mathbf{b} \in \mathbb{R}^{128}$ are a learnable projection and bias mapping each clip embedding into a 128-dimensional attention space, $\mathbf{v} \in \mathbb{R}^{128}$ is a learnable context vector that scores each projected clip, $\tanh(\cdot)$ is applied element-wise, m_t and $m_{t'}$ mask out padded clips before normalization, $\alpha_t \in [0, 1]$ is the resulting attention weight for clip t (with $\sum_{t \in \mathcal{S}} \alpha_t = 1$), and $\mathbf{p}_{\text{attn}} \in \mathbb{R}^{256}$ is the attention-weighted pooled representation. *Transformer Encoder.* $\mathbf{z}_t$ is projected to a 64-dim bottleneck with a rectified linear unit (ReLU) nonlinearity and dropout 0.5, then a single-layer, 4-head Transformer encoder (dropout 0.5) processes the sequence with a prepended learnable classification (CLS) token, as in BERT; the CLS output is the pooled representation, with padding excluded via a key-padding mask. *Bidirectional GRU with Attention.* A two-layer bidirectional gated recurrent unit (GRU; hidden size 128 per direction, giving a 256-dim output after concatenation) scans the clip sequence forward and backward, followed by the attention of Eq. (2) applied to the GRU outputs (in place of $\mathbf{z}_t$). *NetVLAD.* $\mathbf{z}_t$ is projected to a 64-dim bottleneck, then softly assigned to $K = 2$ learnable cluster centroids $\{\boldsymbol{\mu}_k\}_{k=1}^K$, $\boldsymbol{\mu}_k \in \mathbb{R}^{64}$; per-cluster residuals $\sum_{t \in \mathcal{S}} \bar{a}_k(\mathbf{z}_t)(\mathbf{z}_t - \boldsymbol{\mu}_k)$, where $\bar{a}_k(\mathbf{z}_t) \in [0, 1]$ is the soft assignment weight of clip t to cluster k (softmax over the K centroid similarities), are intra-normalized (each cluster's residual vector is L2-normalized independently), concatenated across the $K = 2$ clusters, and L2-normalized as a whole,

TABLE I
MEAN METRICS BY ARCHITECTURE, AVERAGED OVER 6 BACKBONES ($n = 6$ PER ROW; TEST SPEAKERS PER CORPUS: E-DAIC $n = 23$, MODMA $n = 10$)

Set	Architecture	Acc	F1	AUC	Sens	Spec	U
E-DAIC	Mean Pooling	**0.630**	**0.534**	0.613	0.750	0.588	0.696
	Statistical Pool.	0.572	0.511	0.655	0.806	0.490	**0.700**
	Self-Attention	0.551	0.507	0.598	0.778	0.471	0.675
	Bi-GRU + Attn	0.601	0.508	**0.662**	0.722	**0.559**	0.668
	NetVLAD	0.413	0.470	0.607	**0.889**	0.245	0.674
	Transformer Enc.	0.529	0.290	0.577	0.556	0.520	0.544
MODMA	Mean Pooling	0.717	0.745	0.647	0.767	0.667	0.733
	Statistical Pool.	0.650	0.708	0.537	0.800	0.500	0.700
	Self-Attention	0.717	0.731	0.680	0.733	0.700	0.722
	Bi-GRU + Attn	**0.800**	**0.811**	**0.713**	0.833	**0.767**	**0.811**
	NetVLAD	0.733	0.632	0.677	0.567	**0.900**	0.678
	Transformer Enc.	0.517	0.667	0.433	**0.967**	0.067	0.667

Acc = accuracy; F1 = macro or positive-class F1 on the depressed/MDD label; AUC = ROC-AUC; Sens/Spec = sensitivity/specificity as defined in Section III-G; U = clinical utility score of Eq. (3). Each row is the mean over the 6 backbones of Appendix Table IV, one single-seed run per backbone; bold marks the best value in each column within a corpus block.

following [24]. Every architecture ends in a small multi-layer perceptron (MLP) classifier (dropout, one ReLU hidden layer, two-way softmax) that outputs the depressed-vs-healthy decision.

F. Speaker-Independent Splitting

SSL representations encode substantial speaker-identity information alongside paralinguistic content, so leaking clips from one speaker across the training, validation, and test subsets could let a classifier learn to recognize individual voices rather than depression markers, silently inflating reported performance. We therefore use `StratifiedGroupKFold` [29], a splitting procedure that groups all instances belonging to the same entity (here, participant ID) so they cannot be separated across folds, while additionally stratifying by depression label so each split keeps a similar depressed/healthy ratio. This produces a 60/20/20 train/validation/test partition in which every clip from a given participant falls entirely into one split, matching recommended practice for clinical speech datasets [27]. The training split is used to fit model weights, the validation split only to select the best training checkpoint (Section III-G), and the test split is held out and evaluated exactly once per configuration.

G. Training and Evaluation

All 72 configurations are trained independently with AdamW (the Adam optimizer with decoupled weight decay; learning rate 10^{-3}, weight decay 10^{-4}), a class-weighted cross-entropy loss that up-weights the minority class to counter label imbalance, and a fixed random seed for 40 epochs. For each configuration we track the best-F1 epoch and report accuracy, F1, area under the receiver-operating-characteristic curve (ROC-AUC), sensitivity (recall of the depressed/MDD class), and specificity (recall of the healthy-control/HC class), defined as

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \qquad \text{Specificity} = \frac{TN}{TN + FP},$$

TABLE II
MEAN METRICS BY BACKBONE, AVERAGED OVER 6 ARCHITECTURES ($n = 6$ PER ROW; COLL. = NUMBER OF COLLAPSED CONFIGURATIONS OUT OF 6 ARCHITECTURE PAIRINGS)

Set	Backbone	Acc	F1	AUC	Sens	Spec	U	Coll.
E-DAIC	WavLM-Base-Plus	**0.775**	**0.601**	0.676	0.611	**0.833**	0.685	0/6
	WavLM-Large	0.710	0.465	0.673	0.583	0.755	0.641	1/6
	HuBERT-Large	0.522	0.403	0.582	0.722	0.451	0.632	2/6
	Wav2Vec2-Robust	0.268	0.416	0.543	**1.000**	0.010	0.670	**5/6**
	Data2Vec-Audio-L	0.522	0.451	0.611	0.722	0.451	0.632	2/6
	XLS-R-1B	0.500	0.484	**0.626**	0.861	0.373	**0.698**	2/6
MODMA	WavLM-Base-Plus	0.767	0.759	**0.687**	0.700	**0.833**	0.744	1/6
	WavLM-Large	0.717	0.740	0.640	0.767	0.667	0.733	1/6
	HuBERT-Large	**0.800**	**0.806**	0.700	0.767	0.833	**0.789**	1/6
	Wav2Vec2-Robust	0.517	0.564	0.553	0.833	0.200	0.622	**5/6**
	Data2Vec-Audio-L	0.600	0.696	0.507	**0.900**	0.300	0.700	3/6
	XLS-R-1B	0.733	0.731	0.600	0.700	0.767	0.722	1/6

Columns as in Table I; each row is the mean over the 6 architectures of Section III-E for a fixed backbone. Coll. column matches the per-backbone bars of Fig. 3.

, where TP, FN, TN, FP are the counts of true positives, false negatives, true negatives, and false positives on the test-set speaker predictions (positive = depressed/MDD), plus a clinical utility score

$$U = \frac{2 \cdot \text{Sensitivity} + \text{Specificity}}{3} \tag{3}$$

, where $U \in [0, 1]$ (higher is better) weights sensitivity twice as heavily as specificity to reflect the asymmetric cost of a missed diagnosis versus a false alarm. We flag a *collapsed* run as one with (sensitivity, specificity) $= (1, 0)$ or $(0, 1)$, i.e., the model emits the same label for every test speaker.

IV. EXPERIMENTAL RESULTS

Fig. 1 summarizes the pipeline. We report aggregated performance by architecture and by backbone, the best single configuration per corpus, the collapse phenomenon, and seed sensitivity on a focal subset.

A. Aggregated Performance by Architecture and Backbone

Table I averages each architecture's metrics across all six backbones, per corpus. On E-DAIC no architecture wins outright: Mean Pooling has the highest accuracy and F1, but Statistical Pooling and NetVLAD reach higher sensitivity and utility, while NetVLAD's low specificity (0.245) shows it is largely over-predicting the depressed class. On MODMA the picture is cleaner: Bi-GRU with Attention leads on every metric, averaged across all six backbones.

Table II takes the complementary view: metrics per SSL backbone, averaged over all six architectures. Denoting a metric $M(a, b)$ for architecture a on backbone b, Table I reports $\frac{1}{6}\sum_b M(a, b)$ (average over the 6 backbones b, for fixed architecture a) and Table II reports $\frac{1}{6}\sum_a M(a, b)$ (average over the 6 architectures a, for fixed backbone b). The two views agree on the headline conclusion: WavLM-Base-Plus is the strongest, most stable E-DAIC backbone (mean F1 0.601, 0/6 collapsed), while HuBERT-Large edges it out on MODMA (mean F1 0.806 vs. 0.759). Wav2Vec2-Robust is the weakest and least stable backbone on both corpora, collapsing on 5 of

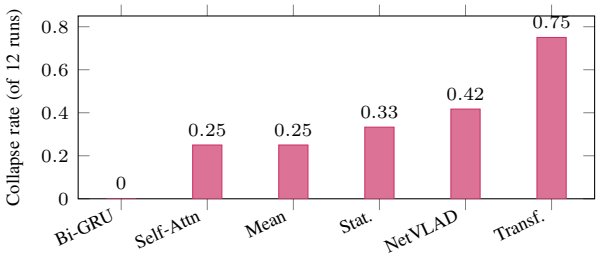

Fig. 2. Share of runs (of 12: 6 backbones × 2 corpora) in which an architecture collapses to a single class, under single-seed training. Bi-GRU with Attention never collapses here, but Section IV-D shows this does not survive seed replication.

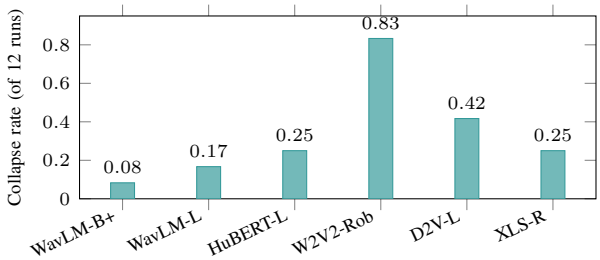

Fig. 3. Share of runs (of 12: 6 architectures × 2 corpora) in which a given backbone collapses to a single class, computed from the **Coll.** column of Table II. Backbone identity predicts collapse at least as strongly as aggregation architecture (Fig. 2).

its 6 architecture pairings per corpus: the clearest single signal that backbone choice can dominate architecture choice.

B. Best Single Configurations

On the E-DAIC, the best configuration by F1 is Self-Attention pooling on WavLM-Base-Plus (accuracy 0.870, F1 0.667, sensitivity 0.500, specificity 1.000). On the MODMA, it is Bi-GRU with Attention on Data2Vec-Audio-Large (accuracy 0.900, F1 0.909, sensitivity 1.000, specificity 0.800), a stronger result across all axes, and the architecture is already flagged as most consistent in Table I.

C. The Collapse Phenomenon

Of the 72 configurations, **24 (33%)** collapse to a single predicted class for every test speaker, unevenly spread across architectures (Fig. 2): the Transformer encoder collapses on 9/12 runs (75%), NetVLAD on 5/12 (42%), while Mean Pooling and Self-Attention each collapse on 3/12 (25%) and Statistical Pooling on 4/12 (33%). Bi-GRU with Attention is the only architecture that never collapses across these single-seed runs (0/12); Section IV-D shows this apparent stability is itself seed-dependent.

Collapse is at least as strongly tied to backbone as to architecture. Wav2Vec2-Robust collapses on 10/12 runs (83%) regardless of pooling head, Data2Vec-Audio-Large on 5/12 (42%), and WavLM-Base-Plus is the most stable at 1/12 (8%), as summarized in Fig. 3. A benchmark testing only Wav2Vec2-Robust would conclude aggregation for depression detection barely works; one testing only WavLM-Base-Plus would miss the collapse phenomenon almost entirely.

Both conclusions would be artifacts of the one backbone chosen. Reading Table I and Table II together, neither factor alone predicts outcome: Bi-GRU with Attention stays in a comparatively high, low-variance band across backbones, while the Transformer encoder swings from competitive scores to collapse; it is the *combination* of architecture and backbone that determines whether a configuration works (full per-cell grid released with the code).

D. Seed Sensitivity: A Focal Subset

The single-seed grid singles out Bi-GRU with Attention as the only architecture that never collapses, but is that stability itself stable? We re-ran the most and least stable architecture/backbone pairs (Bi-GRU with Attention and the Transformer encoder, on WavLM-Base-Plus and Wav2Vec2-Robust) across three random seeds on both corpora (Table III). The apparent stability of Bi-GRU does not survive replication: on MODMA with WavLM-Base-Plus it averages only F1 0.31 ± 0.42 at sensitivity 0.25, collapsing toward the healthy class on some seeds, the very failure mode it never exhibited in the single-seed grid. Standard deviations reach 0.42 in F1, and no configuration sustains a balanced trade-off once seeds vary; single-seed evaluation understates instability just as single-backbone evaluation does.

V. Discussion and Analyses

A. Robustness vs. average performance

A benchmark reporting only mean F1 per architecture (Table I) would hide that a third of the underlying runs never learn to discriminate at all. Two architectures can have similar means for different reasons: one by reliably landing in a moderate range, another by alternating between strong runs and total collapse. Under single-seed training, only Bi-GRU with Attention never fails outright, but the replication in Section IV-D shows even that reliability is fragile: no head can be recommended as a default until robustness is established across both backbones and seeds.

B. Why do some architectures collapse more?

After merging the train and validation, E-DAIC has 99 training participants, and MODMA has 42. The Transformer encoder and NetVLAD both introduce a routing mechanism (CLS-token self-attention; soft cluster assignment) that must be learned essentially from scratch on top of a frozen backbone with only a few dozen training sequences. Under class-weighted cross-entropy, an under-trained routing function can minimize loss fastest by degenerating to a fixed-label rule, which is what we observe. This pattern is consistent with the broader neural-collapse literature, where late-training features and classifier weights for small, imbalanced problems converge onto a degenerate geometry that a majority-class rule already satisfies [28]. The bidirectional GRU instead updates its hidden state incrementally at every clip, which may give it a smoother optimization landscape in the same low-data regime, though the seed-replication subset shows it, too, degenerates on some seeds, so the effect is one of degree rather than a categorical difference; confirming this mechanism directly, for

TABLE III

SEED REPLICATION ON THE FOCAL SUBSET (MEAN±STD OVER 3 SEEDS $\{0, 1, 2\}$; FINAL ROW SINGLE-SEED ONLY). CORPUS/BACKBONE/ARCH. PAIRS WERE CHOSEN AS THE MOST STABLE (BI-GRU, WAVLM-B+) AND LEAST STABLE (TRANSF., W2V2-ROB) CELLS OF FIG. 2/3.

Corpus	Backbone	Arch.	F1	Sens	Spec
E-DAIC	WavLM-B+	Bi-GRU	0.458±0.115	0.667	0.515
E-DAIC	WavLM-B+	Transf.	0.432±0.027	0.792	0.324
E-DAIC	W2V2-Rob	Bi-GRU	0.388±0.027	0.625	0.456
E-DAIC	W2V2-Rob	Transf.	0.406±0.016	0.917	0.088
MODMA	WavLM-B+	Bi-GRU	0.306±0.419	0.250	1.000
MODMA	WavLM-B+	Transf.	0.250±0.319	0.300	0.750
MODMA	W2V2-Rob	Bi-GRU	0.679±0.024	1.000	0.050
MODMA	W2V2-Rob	Transf.	0.667	1.000	0.000

F1/Sens/Spec defined as in Section III-G; Sens/Spec columns show the mean over 3 seeds (per-seed values in released results). MD-W2-T (last row) was run for a single seed only, so no standard deviation is reported.

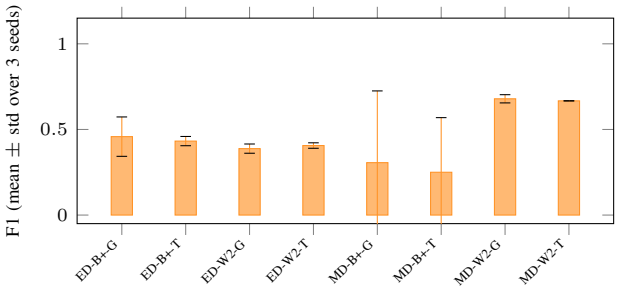

Fig. 4. Seed replication on the focal subset of Table III (ED = E-DAIC, MD = MODMA, B+ = WavLM-Base-Plus, W2 = Wav2Vec2-Robust, G = Bi-GRU with Attention, T = Transformer encoder). Error bars are one standard deviation over 3 seeds (MD-W2-T is a single seed). The widest bars, both on MODMA with Bi-GRU, show that single-seed stability does not imply multi-seed stability.

instance by tracking embedding variance or gradient norms during training, is beyond what our benchmark alone can establish and is left to future work.

C. Backbone choice is not a side detail

Wav2Vec2-Robust's 83% collapse rate is the largest single effect in the grid, larger than the gap between any two architectures. Its pretraining emphasizes invariance to acoustic domain shift (noise, channel) rather than the paralinguistic detail WavLM's utterance-mixing objective seems to preserve [4], [8]; because our backbone set mixes self-supervised-only and ASR-fine-tuned checkpoints (Appendix Table IV), this comparison is suggestive rather than fully controlled. A study reporting a strong or weak aggregation architecture without also reporting its backbone gives only half the answer.

D. MODMA vs. E-DAIC

The best achievable F1 is consistently higher on MODMA than E-DAIC (0.909 vs. 0.667). MODMA's clinician-based diagnoses on a constrained protocol plausibly make the classification problem itself easier than E-DAIC's noisier PHQ-8 self-report labels and virtual-agent format. We would caution against reading MODMA's numbers as evidence the problem is close to solved: its test set has only 10 speakers, so one misclassified participant moves accuracy by 10 points.

VI. LIMITATIONS AND ETHICAL CONSIDERATIONS

The 72-cell grid uses a single fixed seed. This prevents per-cell metrics from separating architectural effects from random noise. Our replication shows this matters as F1 standard deviation reaches 0.42 and the Bi-GRU stability result does not hold across seeds. Backbone effects appear more robust than architecture rankings. A multi-seed sweep remains for future work. Due to computational costs, checkpoint selection tracks best-F1 on the held-out test partition rather than a validation criterion. Absolute metrics likely overstate generalization, though comparative claims remain robust since model collapse is a coarse failure. Small test sets (23 E-DAIC and 10 MODMA speakers) make metrics sensitive to individual participants. Additionally, mean-pooling may cap performance across all architectures, and the optimization-landscape hypothesis requires further verification.

This work uses two existing de-identified corpora under established agreements. DEPOOL is a research benchmark rather than a diagnostic tool. E-DAIC labels reflect binarized PHQ-8 self-reports, and MODMA labels do not support clinical deployment. Known demographic imbalances in these corpora preclude claims of fairness, and future work must prioritize subgroup evaluation. Real-world applications would require prospective clinical validation, informed consent, and human oversight.

VII. CONCLUSIONS

We presented DEPOOL, which combines six temporal aggregation architectures with six frozen SSL speech backbones and two depression corpora, yielding 72 controlled configurations; layer selection is handled by a learned softmax rather than a hand-picked layer. The central finding is that aggregation architecture cannot be evaluated in isolation from its backbone or random seed: a third of all 72 single-seed configurations collapse into trivial single-class prediction, concentrated in specific architecture/backbone combinations (the Transformer encoder and NetVLAD; the Wav2Vec2-Robust backbone) rather than being evenly distributed. Bidirectional GRU with attention is the one architecture that never collapses in the single-seed grid, but seed replication on a focal subset breaks even that stability, so we deliberately stop short of recommending any head as a reliable default. The single best configuration, Bi-GRU with Attention on Data2Vec-Audio-Large on MODMA, reaches F1 = 0.909, but no configuration reaches that reliability on the harder E-DAIC corpus, and seed replication cautions that such best-case numbers are optimistic: gains here are corpus-, backbone-, and seed-dependent at once. Future work should extend multi-seed replication to all 72 cells, adopt a strict validation-only checkpointing criterion, and extend the grid to additional backbones, languages, and within-clip temporal encodings.

REFERENCES

[1] World Health Organization, "Depressive disorder (depression)," WHO Fact Sheet, Sep. 2023.

[2] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[3] D. M. Low, D. Bentley, and S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investigative Otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[4] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Li, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.

[5] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, "The distress analysis interview corpus of human and computer interviews," in *Proc. LREC 2014*, 2014.

[6] H. Cai, Z. Yuan, Y. Gao, S. Sun, N. Li, F. Tian, H. Xiao, J. Li, Z. Yang, X. Li, Q. Zhu, Z. Xiang, and N. Sun, "MODMA dataset: A multi-modal open dataset for mental-disorder analysis," *arXiv preprint arXiv:2002.09283*, 2020.

[7] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *arXiv preprint arXiv:2106.07447*, 2021.

[8] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," in *Proc. Interspeech 2021*, pp. 721–725, 2021.

[9] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proc. ICML 2022*, pp. 1298–1312, 2022.

[10] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[11] B. Maji, R. Guha, A. Routray, S. Nasreen, and D. Majumdar, "Investigation of layer-wise speech representations in self-supervised learning models: A cross-lingual study in detecting depression," in *Proc. Interspeech 2024*, pp. 3020–3024, 2024.

[12] S.-W. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Qian, P. Hira, and J. Glass, "SUPERB: Speech processing universal performance benchmark," in *Proc. Interspeech 2021*, pp. 1194–1198, 2021.

[13] H. Wu, H.-C. Chou, K.-W. Chang, L. Goncalves, J. Du, J.-S. R. Jang, C.-C. Lee, and H.-y. Lee, "EMO-SUPERB: An in-depth look at speech emotion recognition," *arXiv preprint arXiv:2402.13018*, 2024.

[14] Z. Han, T. Geng, H. Feng, J. Yuan, K. Richmond, and Y. Li, "Cross-lingual speech emotion recognition: Humans vs. self-supervised models," *arXiv preprint arXiv:2409.16920*, 2024.

[15] L. Albuquerque, A. R. S. Valente, A. Teixeira, D. Figueiredo, P. Sa-Couto, and C. Oliveira, "Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan," *PLOS ONE*, vol. 16, no. 4, p. e0248842, 2021.

[16] T. Alhanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," in *Proc. Interspeech 2018*, pp. 1716–1720, 2018.

[17] P. Zhang, M. Wu, H. Dinkel, and K. Yu, "DEPA: Self-supervised audio embedding for depression detection," in *Proc. ACM MM 2021*, pp. 135–143, 2021.

[18] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.

[19] J. R. Williamson, T. F. Quatieri, B. S. Helfer, J. Perricone, S. S. Ghosh, and G. Ciccarelli, "Tracking depression-related mental state using multimodal analysis," in *Proc. ACM ICMI 2013*, pp. 279–286, 2013.

[20] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, and M. Pantic, "AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition," in *Proc. ACM MM AVEC 2019*, pp. 3–12, 2019.

[21] Y. F. A. Gaus, S. Ballard, N. Cummins, and B. Schuller, "Speaker normalization for speech-based depression detection," in *Proc. ICASSP 2021*, pp. 7278–7282, 2021.

[22] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "DepAudioNet: An efficient deep model for audio based depression classification," in *Proc. ACM AVEC 2016*, pp. 35–42, 2016.

[23] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. ICASSP 2017*, pp. 2227–2231, 2017.

[24] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. CVPR 2016*, pp. 5297–5307, 2016.

[25] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. Interspeech 2018*, pp. 2252–2256, 2018.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL 2019*, pp. 4171–4186, 2019.

[27] B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen, S. Ottl, M. Gerczuk, P. Tzirakis, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, et al., "The INTERSPEECH 2021 computational paralinguistics challenge," in *Proc. Interspeech 2021*, pp. 431–435, 2021.

[28] V. Papyan, X. Y. Han, and D. L. Donoho, "Prevalence of neural collapse during the terminal phase of deep learning training," *Proc. Natl. Acad. Sci.*, vol. 117, no. 40, pp. 24652–24663, 2020.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

APPENDIX

Table IV summarizes the six self-supervised learning (SSL) speech backbones evaluated within the DEPOOL benchmark. We select these models to represent diverse pretraining objectives, including contrastive learning, masked prediction, and self-distillation, as well as varying parameter counts and training data sizes. The backbones range from the 94M-parameter WavLM-Base-Plus to the 965M-parameter XLS-R-1B, encompassing both purely self-supervised models and those further fine-tuned for automatic speech recognition (ASR-FT). For each backbone, we provide the number of available Transformer hidden-state layers, L, and the hidden dimension, D, both of which are used by the semi-fine-tuned layer-aggregation protocol defined in Section III-D. All backbones are kept frozen throughout our training process, ensuring that the benchmark results reflect the downstream aggregation architecture and the learned layer-weighting efficacy rather than backbone-internal weight updates.

TABLE IV
SSL BACKBONES EVALUATED IN DEPOOL.

Backbone	**Layers**	**Dim.**	**Params**	**Pretrain.**	**Ref.**
WavLM-Base-Plus	13	768	94M	SSL, 94k h Libri-Light/GigaSpeech/VoxPopuli	[4]
WavLM-Large	25	1024	316M	SSL, 94k h Libri-Light/GigaSpeech/VoxPopuli	[4]
HuBERT-Large	25	1024	316M	SSL on 60k h Libri-Light, then ASR-FT on 960 h LS	[7]
Wav2Vec2-Robust	25	1024	317M	SSL, domain-mixed (read + noisy + telephone speech)	[8]
Data2Vec-Audio-Large	25	1024	314M	SSL on 60k h Libri-Light, then ASR-FT on 960 h LS	[9]
XLS-R-1B	49	1280	965M	SSL, 436k h multilingual (128 languages)	[10]

SSL = self-supervised only; SSL+FT = self-supervised then ASR fine-tuned (HuBERT-Large: `hubert-large-ls960-ft`; Data2Vec-Audio-Large: `data2vec-audio-large-960h`). Layers = number of Transformer hidden-state outputs, including the feature-extractor output, available to the featurizer of Eq. (1). Dim. = hidden size D of each layer. Params = approximate total parameter count of the released checkpoint (frozen; not updated during our training). LS = LibriSpeech.