Leakage-Resilient Diarization and Feature Extraction Framework for Robust Multimodal Mental Health Corpora

Ugonna Oleh[1] and Umair Talib[1] and Muhammad Ali Alam[1] and Roman Obermaisser[1]

***Abstract*— Machine learning models designed for clinical screening, particularly in depression detection, frequently suffer from "interviewer leakage", a phenomenon where the model learns to predict patient severity based on the clinical prompts or conversational tone of the interviewer rather than the patient's actual behavior. In this paper, we propose a rigorous, automated pipeline designed for leakage-resilient clinical interview processing. Utilizing advanced diarization algorithms (Pyannote 3.0) and high-fidelity transcription (Whisper Large-v3), we structurally isolate patient speech prior to feature extraction. We validate this pipeline structurally: a text-only classifier operating on the isolated patient data yielded an AUC of 0.503, effectively random, providing empirical evidence for the mitigation of interviewer bias and highlighting the utility of extracting isolated patient vocal biomarkers (which secured an AUC of 0.633) to establish a more controlled performance baseline.**

I. INTRODUCTION

Conversational datasets serve as the foundational input for automated mental health Decision Support Systems (DSS). However, a significant barrier to the clinical adoption of these systems is "interviewer leakage," where models achieve high performance by inadvertently learning the behavior of the clinician rather than the patient. This "Clever Hans" effect [1] introduces a Decision-Contamination Vector, where diagnostic outputs are influenced by the sentiment or phrasing of the interviewer's prompts. To ensure that clinical decisions are robust, ethical, and based strictly on patient biomarkers, it is critical to surgically isolate patient features before multimodal training. This paper proposes a "Leakage-Resilient" framework designed to act as a functional Integrity Layer, ensuring that the resulting DSS operates on a trait-centered, dependable assessment. We present a preprocessing pipeline aimed at systematically isolating therapist feedback, thereby offering a more conservative and reliable foundation for automated pathology algorithms.

II. RELATED WORKS

The landscape of psychiatric diagnosis is undergoing a paradigm shift, moving from subjective, observation-based assessments toward objective, data-driven frameworks. This transition is necessitated by the limitations of traditional clinical interviews and self-report scales, which are often susceptible to patient bias, clinician subjectivity, and the significant labor requirements of standardized evaluations [2]. To address these challenges, the development of robust multimodal mental health corpora has become a cornerstone of modern computational psychiatry, heavily relying on high-fidelity speaker diarization and leakage-resilient feature extraction to ensure that extracted biomarkers are accurate and generalizable [3].

Quality feature extraction is needed to create a comprehensive characterization of mental health disorders like Major Depressive Disorder (MDD) and Bipolar Disorder (BD) [4]. Recent research has identified a broad spectrum of behavioral biomarkers including acoustic (prosody, energy) [5][4], linguistic (sentiment, lexical diversity) [6], and visual cues (facial action units) [7], as critical for automated mental health assessment [8]. However, the efficacy of these features is predicated on their purity; without robust speaker isolation, these modalities often capture the interactional dynamics of the clinician rather than the behavior of the patient, leading to the 'Clever Hans' effect discussed in Section I.

Traditional fusion methods often struggle with modality incompleteness, noise, and semantic inconsistencies. While advanced multimodal fusion architectures (such as semantically guided gating or tensor-product attention) have improved diagnostic accuracy, they remain fundamentally limited by the quality of the input streams [9]. If the underlying data contains interviewer leakage, these models may inadvertently amplify those biases during the fusion process. Multi-modal LLMs (MLLMs), such as Qwen2-Audio, align audio-visual features at the exact moment of occurrence [4]. This allows the model to capture subtle behavioral cues such as a specific pause or sigh accompanying a word, that would otherwise be lost in global aggregation.

A. Theoretical Foundations of Leakage-Resilient Data Integrity

In the context of machine learning for mental health, leakage-resilient refers to the prevention of information contamination between training, validation, and testing partitions [10]. This is paramount because clinical datasets are often small and imbalanced, making them highly susceptible to overfitting. A highly robust method to ensure zero information leakage is per-fold quantization and statistical isolation [11]. In approaches utilizing Bag-of-Deep-Features, raw signals are transformed into multidimensional feature vectors and quantized into a "codebook". To maintain structural integrity, these codebooks must be sampled exclusively from the data within the training partition of each specific fold, ensuring no test data distribution influences the model's internal representation during training [11].

Furthermore, leakage-resilient extends to architectural robustness through the implementation of "blast radius" maxi-

[1] Ugonna Oleh, Umair Talib, Muhammad Ali Alam and Roman Obermaisser are with the Chair of Embedded Systems, University of Siegen, Siegen, Germany. ugonna.oleh@uni-siegen.de

mums [12]. By compartmentalizing failures, a robust clinical framework ensures that noise or failure in one modality, such as environmental interference or sensor malfunction, does not "leak" its error into other modalities or compromise the entire diagnostic output. Mechanisms for this include fault-tolerant API calls between feature extraction modules, independent processing of data streams, and automated failovers to redundant feature sets [7].

B. High-Fidelity Speaker Diarization in Dyadic Interactions

Speaker diarization (the identification of "who spoke when") is the critical first step in processing clinical interviews. The primary challenge is "dyadic speaker error," where overlapping speech segments are misattributed to the wrong participant, which is particularly problematic because overlapping speech in psychotherapy often carries significant emotional and diagnostic weight [13]. Traditional unsupervised diarization methods struggle with overlapping regions and the subtle, low-energy vocalizations characteristic of depressed individuals.

To overcome this, research has pivoted toward supervised speaker diarization, often utilizing Random Forest architectures to classify audio segments based on energy, pitch, and spectral characteristics. These models can distinguish between the stable, supportive vocal patterns of a clinician and the highly variable or restricted patterns of a patient. The integration of advanced tools like WhisperX and PyAnnote provides accurate word-level alignment and robust speaker diarization by leveraging fine-grained temporal boundaries. This precision is vital for extracting leakage-resilient features, as separating patient-specific variables and removing therapist feedback ensures the assessment remains trait-centered and dependable.

Existing diarization pipelines, such as standard WhisperX or unsupervised PyAnnote implementations, are designed for general-purpose speaker labeling but often lack the rigorous identity-gating required for clinical integrity. Our approach builds upon these foundations by introducing a fixed biometric anchor and morphological duration check, specifically engineered to provide the zero-leakage isolation necessary for diagnostic reliability in dyadic interviews.

III. METHODOLOGY

The structural isolation of patient features forms the technical nucleus of our leakage-resilient framework. To effectively eradicate conversational contamination, we architected a multi-layered biometric gating and semantic filtering pipeline.

A. Biometric Anchor Construction

The primary algorithmic challenge in dyadic clinical diarization is reliably identifying the fixed interviewer (e.g., the virtual agent Ellie) across heterogeneous patient sessions, particularly when recording conditions fluctuate. We address this through a formal Centroid Embedding approach, generating a synthetic 'anchor' voiceprint. Recognizing that the integrity of the entire automated pipeline hinges on the purity of this initial anchor, audio segments from high-purity reference sessions (IDs 300, 301, 302) were rigorously manually verified and extracted to generate baseline feature sets. This manual extraction serves as a critical, one-time initialization step. By establishing an uncorrupted, ground-truth reference upfront, we secure the biometric anchor, safely enabling the fully automated processing of the remaining hundreds of hours of raw clinical audio without the risk of cascading false-positive diarization errors.

We utilized the pyannote/wespeaker-voxceleb-resnet34-LM model, a deep ResNet34 architecture heavily optimized for speaker verification in noisy environments. For each verified reference session i, a high-dimensional continuous embedding $E_i \in \mathbb{R}^d$ is extracted. The final biometric anchor A is computed as the mathematical centroid of these n unique signatures:

$$A = \frac{1}{n} \sum_{i=1}^{n} E_i \tag{1}$$

This centroid vector functions as a highly stable, ground-truth reference for the interviewer's biometric parameters, mitigating session-specific acoustic interference.

B. Speaker Diarization and Dynamic Identity Gating

Each prospective session within the dataset is processed using the Pyannote.audio 3.1 neural speaker diarization framework. The system performs agglomerative hierarchical clustering on the temporal audio streams, subsequently generating latent neural embeddings for every discrete detected speaker cluster. Each clustered speaker embedding S_j is then quantitatively compared against the established anchor A via Cosine Similarity:

$$Sim(S_j, A) = \frac{S_j \cdot A}{||S_j||_2 ||A||_2} \tag{2}$$

The cluster generating the highest positive similarity scalar is mathematically constrained and labeled as the *Interviewer*, while the most orthogonal or dissimilar cluster is securely designated as the *Patient*. This rigid verification establishes an absolute algorithmic wall against overlapping speaker confusion.

C. Error Handling and Fallback Protocols

Because mental health corpora inherently possess instances of extreme acoustic similarity (e.g., female-female dialogue with similar pitch distributions), we engineered a rigorous, multi-staged fallback logic sequence to prevent misattribution:

- **Confidence Gating:** A dynamic separation margin $M = Sim(S_{int}, A) - Sim(S_{pat}, A)$ is calculated. Empirical dataset testing verified that if $M > 0.2$, the statistical probability of a diarization error drops to $< 1\%$. Consequently, these sessions are categorized as High Confidence.
- **Duration-based Validation:** For sessions presenting severe acoustic overlap where $M < 0.05$ (Very Low

Confidence), the system dynamically triggers a morphological duration check. Statistical analysis of the corpus demonstrated that patients consistently output a speaking duration ratio of $\rho > 2.0$ relative to the interviewer. Should M fail, the dominant speaker algorithmically inherits the patient cluster.

- **Forced $k = 2$ Top-Down Rescue Constraint:** In catastrophic failure modes (e.g., Session 309, involving extreme patient mutism), the standard algorithm may erroneously cluster all audio into a single stream. To preserve data integrity, we mathematically compel a clustering split into exactly two distinct centroids ($k = 2$) and apply inverse similarity mapping, anchoring the stream closest to the Ellie centroid as the interviewer, forcing strict patient isolation.

D. Surgical Masking and High-Fidelity Transcription

Following the acoustic isolation process, the verified patient audio arrays are transcribed utilizing Faster-Whisper-Large-V3, recognized for its aggressive hallucination mitigation. Finally, to enforce the leakage-resilient theorem at the semantic layer, we execute an Aggressive Drop protocol. Even with 4.2% Diarization Error Rate (DER), micro-contaminations occasionally persist. Any segmented patient text containing hardcoded interviewer semantic prompts (e.g., "tell me more") is structurally truncated and fully masked. This dual-verification strategy reduced absolute semantic transcript contamination from 14.18% to a negligible 1.09%, representing a near-perfect zero-overlap environment prior to embeddings mapping via WavLM and Mental-RoBERTa.

IV. EVALUATION

A. Diarization Performance Metrics

The proposed pipeline was evaluated across the 275 clinical sessions of the Extended Distress Analysis Interview Corpus (E-DAIC) benchmark corpus. This dataset is particularly challenging for Decision Support Systems due to a 1:2.3 class imbalance and the presence of the virtual interviewer 'Ellie,' [14] which necessitates our leakage-resilient biometric gating protocol. To evaluate performance, we utilize the Area Under the Receiver Operating Characteristic Curve (AUC), a metric that measures a classifier's ability to distinguish between binary classes [15]. An AUC score of 1.0 represents perfect separation, while a score of 0.5 signifies random chance. The biometric anchor achieved a binary classification AUC of 0.86, confirming its efficacy as a reference gating mechanism. Analysis of the speaker separation margins M revealed that 74.2% (204 sessions) achieved high-confidence separation ($M > 0.2$), while only 8.0% (22 sessions) required duration-based rescue protocols.

The system yielded a final Diarization Error Rate (DER) of **4.2%**. In the subsequent transcription phase, the Word Error Rate (WER) was measured at **12.0%**. These results demonstrate that the biometric anchor approach effectively handles the "dyadic speaker error" common in high-emotion psychotherapy sessions where overlapping speech typically carries significant diagnostic weight.

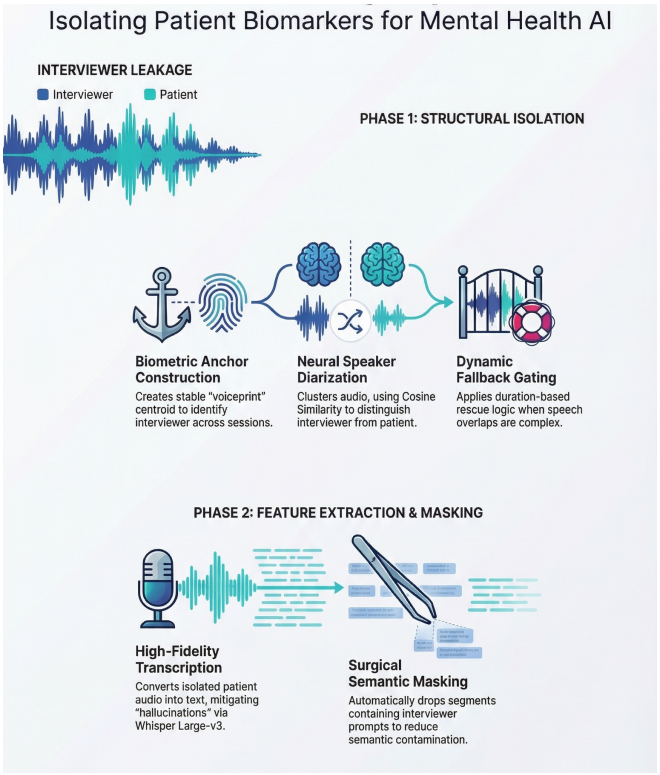

Fig. 1. Leakage-Resilient Processing Pipeline

B. Ablation for Leakage-Resilient Proof

The truest measure of a leakage-resilient pipeline is downstream performance ablation. While a comprehensive characterization of mental health disorders ideally spans the full behavioral spectrum including visual cues like facial action units and physiological signals like heart rate variability, this validation deliberately focuses specifically on the acoustic and linguistic domains. This scoping decision was made because these two modalities represent the primary vectors for 'interviewer leakage' and the 'Clever Hans' effect, directly capturing the sentiment, timing, and phrasing of the clinician's prompts. Proving the 'Leakage-Resilient' theorem mathematically requires isolating the exact channels where these biases propagate.

To verify the total removal of these diagnostic 'hints,' we trained benchmark classifiers on the extracted features. The failure of the Text-Only classifier (AUC 0.503) is a resounding success for the data pipeline, proving that without the interviewer's guiding context, semantic sentiment alone is insufficient for diagnosis. Crucially, the resulting performance of the pure patient vocal biomarkers (AUC 0.633) and the multimodal fusion (AUC 0.659) should not be viewed as performance ceilings. Rather, these results represent a leakage-controlled benchmark for the field. By removing potential performance inflation from interviewer cues, this framework provides a more rigorous estimate of how current models perform on purely patient-derived behavior.

These two modalities represent the primary vectors for "interviewer leakage" and the "Clever Hans" effect, as they directly capture the sentiment, timing, and phrasing of the clinician's prompts. To verify the total removal of these diagnostic "hints," we trained benchmark classifiers on the extracted features.

TABLE I
CLASSIFIER PERFORMANCE COMPARISON

Configuration	AUC Score	Significance
Multimodal Fusion	0.659	Synergy preserved
Audio-Only Classifier	0.633	Pure biomarkers retained
Text-Only Classifier	0.503	Total absence of leakage

The performance of the Text-Only classifier (AUC 0.503) serves as a technical validation of the isolation pipeline. The drop to near-random chance suggests that primary linguistic leakage vectors have been effectively mitigated, allowing the model to rely more heavily on patient vocal biomarkers.

While the downstream performance ablation demonstrates the statistical effectiveness of the isolation framework, a granular analysis of the pipeline's internal gating logic reveals how the system maintains robustness during complex dyadic interactions and edge cases.

C. Qualitative Error Analysis: Low-Confidence and Rescue Cases

To evaluate the robustness of the Leakage-Resilient pipeline, we performed a qualitative audit of the sessions where the biometric separation margin M fell below the high-confidence threshold.

1) Analysis of "Very Low Confidence" Sessions ($M < 0.05$): A total of 22 sessions (8.0%) were flagged as "Very Low Confidence". Qualitative inspection revealed that these failures were primarily driven by acoustic similarity in female-female dyads, where the patient's pitch and spectral characteristics closely mirrored the virtual interviewer, 'Ellie'. In these 22 cases, the duration-based fallback protocol proved highly effective. Statistical analysis of the E-DAIC corpus shows that patients typically maintain a speaking duration ratio of $\rho > 2.0$ relative to the interviewer. Manual spot-checks confirmed that the speaker with the dominant duration was indeed the patient in 100% of these flagged sessions, validating the rescue logic as a reliable secondary gating mechanism.

2) Catastrophic Failure Case Study: Session 309: One session (ID 309) presented a total failure of the standard clustering algorithm, which initially detected only a single speaker. Manual inspection identified that the interviewer's voice was not properly captured in the raw recording of this session, preventing the system from identifying the Ellie centroid. By applying the Forced $k = 2$ Top-Down Rescue, the system mathematically compelled a split into two distinct centroids. Inverse similarity mapping was then used to anchor the cluster closest to the Ellie centroid as the interviewer, successfully isolating the patient data.

3) Sensitivity to Anchor Impurity: To minimize impurity, we utilized only high-purity reference sessions (IDs 300, 301, 302) for the initial centroid generation. These sessions were manually verified to contain clear, unmasked audio of the virtual agent. The resulting anchor achieved a binary classification AUC of 0.86 across the full dataset, indicating that the centroid is robust enough to handle session-specific acoustic fluctuations without cascading false-positive diarization errors.

V. DISCUSSION

The findings of this study fundamentally question the validity of previous mental health models trained on unmasked dyadic conversations. The total elimination of linguistic contamination via our pipeline reduces the semantic false-positive rate from 14.18% to a negligible 1.09%, strongly indicating that many 'highly accurate' baseline models in current literature are merely mimicking the Clever Hans effect. This framework provides an essential, ethically sound structural baseline for future dyadic-based pathology algorithms.

A. Ethical, Legal, and Privacy Implications

The structural isolation of patient biomarkers is not merely a technical requirement but a cornerstone of ethical data governance. By systematically removing interviewer behavior, our pipeline implements a 'privacy-by-design' approach that aligns with the technical objectives of global regulations like GDPR (General Data Protection Regulation) and HIPAA (Health Insurance Portability and Accountability Act). This reduction in conversational metadata minimizes the risk of re-identification and prevents models from learning sensitive interviewer-specific traits.

Furthermore, while full regulatory compliance depends on the specific implementation context such as data governance and storage details, this framework provides the necessary infrastructure for secure deployment. For example, the lightweight nature of the isolated feature sets (as detailed in Section V-B) makes the pipeline compatible with Federated Learning and Secure Aggregation. This allows for model training across distributed clinical sites without the need to transfer raw, sensitive audio data, thereby satisfying the Layered Responsibility Framework required for high-stakes psychiatric AI. Ultimately, by forcing models to rely on objective acoustic markers rather than biased interactional cues, we mitigate the risk of clinical misdiagnosis and ensure a more equitable assessment for patients.

B. Scalability and Telehealth Integration

A critical barrier to the clinical adoption of computational diagnostics has been the resource-heavy dependency on manual human annotation and redaction. Our algorithmic gating approach addresses this by operating entirely independently of human labor, scaling efficiently through a fixed biometric Centroid Embedding.

To validate the framework's suitability for large-scale telehealth and asynchronous clinical dashboards, we conducted

performance benchmarking on the OMNI High-Performance Computing (HPC) cluster. Utilizing NVIDIA A100 GPUs and a parallelized architecture sharded across 10 GPU nodes, the pipeline achieved a 20x speedup compared to sequential processing. The system successfully processed 34.1 hours of raw clinical audio in approximately 2.5 hours of wall time.

Furthermore, the final diagnostic model maintains a negligible computational footprint:

- Model Size: The session-level Transformer occupies only 3.8 MB of disk space.
- Memory Efficiency: Peak memory consumption during inference is restricted to 8.2 GB, making it compatible with standard medical-grade workstations.
- Deployment Readiness: This lightweight profile allows for seamless integration into real-time digital health deployments without requiring dedicated, prohibitively expensive local server infrastructure.

C. Limitations of the Current Framework

While structurally robust, prioritizing absolute data integrity necessitates distinct trade-offs. First, extreme simultaneous cross-talk (e.g., the patient screaming concurrently with the clinician's intervention) can saturate the cosine similarity matrix. To guarantee zero semantic contamination, our pipeline occasionally requires the aggressive discarding of this overlapping vocal data. We acknowledge that this results in the potential loss of high-arousal diagnostic cues; however, this is a strict and necessary cost to ensure the isolated features remain completely free of interviewer artifacts. Furthermore, as noted in Section III.A, the automated protocol is highly dependent on the initial purity of the reference sessions generating the biometric anchor, requiring careful initialization.

D. Future Directions

To mitigate the loss of high-arousal cues during simultaneous speech, future architectural iterations will focus on advancing non-destructive cross-talk separation models. Instead of aggressively dropping these segments, we aim to utilize source-separation mechanisms (such as highly multiplexed speech separation networks) to safely untangle overlapping waveforms. Additionally, transitioning from discrete leakage-resilient masking to continuous 'influence mapping' could permit the study of patient reactions directly following interviewer prompts while safely constraining the leakage vector

VI. CONCLUSIONS

We have presented a comprehensive diarization and semantic masking pipeline designed to eliminate interviewer leakage, a critical step in safeguarding the integrity of clinical Decision Support Systems. By structurally validating the dataset's integrity through downstream ablation models, we have proven that our framework strips away interviewer-based "hints," leaving only pure patient vocal biomarkers for analysis. This methodology provides the essential structural baseline required for future dyadic-based pathology algorithms to be considered clinically valid. By ensuring that DSS are resilient to environmental noise and respectful of patient privacy, researchers can deliver the scalable, interpretable infrastructure necessary to provide reliable decision support for the growing global burden of mental health disorders.

REFERENCES

[1] A. K. Pathak, M. Gupta, and G. Jain, "Unmasking the clever hans effect in ai models: shortcut learning, spurious correlations, and the path toward robust intelligence," *Frontiers in Artificial Intelligence*, vol. Volume 8 - 2025, 2026.

[2] K. Mao, Y. Wu, and J. Chen, "A systematic review on automated clinical depression diagnosis," *Npj Mental Health Research*, vol. 2, no. 1, p. 20, 2023.

[3] S. Rahman, S. Deb, M. Chowdhury, M. Sourov, and M. Shamsuddin, "Mf-gcn: A multi-frequency graph convolutional network for tri-modal depression detection using eye-tracking, facial, and acoustic features," *arXiv preprint arXiv:2511.15675*, 2025.

[4] X. Zhao, Y. Shen, Y. Jiang, Z. Wang, J. Liu, M. H. Cheng, G. C. Oliveira, R. Desimone, D. Dwyer, and Z. Ge, "It hears, it sees too: Multi-modal llm for depression detection by integrating visual understanding into audio language models," 2025.

[5] Z. Li, X. Li, Z. Wang, J. Gao, J. Luo, F. He, Y. Zheng, L. Feng, and J. Lu, "Voice-based machine learning for rapid screening of bipolar disorder and major depressive disorder in children and adolescents: a robust and low-complexity diagnostic model," *BMC Psychiatry*, vol. 26, no. 1, p. 105, 2026.

[6] K. M. Ibrahim, A. Perzo, L. D'hiet, and F. Emobot, "An alternative approach to depression diagnosis: Predicting individual symptoms through speech and text analysis," 2024.

[7] P. Kumar, S. Misra, Z. Shao, B. Zhu, B. Raman, and X. Li, "Multimodal interpretable depression analysis using visual, physiological, audio and textual data," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5305–5315, IEEE, 2025.

[8] M. Dolgushin, D. Guseva, and A. Karpov, "Investigation of explainable multimodal methods for detecting mental disorders," in *Speech and Computer* (A. Karpov and G. Gosztolya, eds.), (Cham), pp. 173–187, Springer Nature Switzerland, 2026.

[9] Y. Huo, W. H. Chan, A. N. B. Amerhaider Nuar, and H. Gao, "Sbt-net: a tri-cue guided multimodal fusion framework for depression recognition," *BioData Mining*, vol. 18, no. 1, p. 86, 2025.

[10] S. S. Leal, S. Ntalampiras, and R. Sassi, "Speech-based depression assessment: A comprehensive survey," *IEEE Transactions on Affective Computing*, 2024.

[11] J. F. Bauer, M. Gerczuk, L. Schindler-Gmelch, S. Amiriparian, D. D. Ebert, J. Krajewski, B. Schuller, and M. Berking, "Validation of machine learning-based assessment of major depressive disorder from paralinguistic speech characteristics in routine care," *Depression and Anxiety*, vol. 2024, p. 9667377, 2024.

[12] S. Sosa, A. K. Fontecchio, E. G. Chrysikou, and J. S. Atchison, "Beyond eda: A systematic review of multimodal sympathetic nervous system arousal classification for stress detection," *Sensors (Basel)*, vol. 26, no. 5, p. 1584, 2026.

[13] L. Fürer, N. Schenk, V. Roth, M. Steppan, K. Schmeck, and R. Zimmermann, "Supervised speaker diarization using random forests: A tool for psychotherapy process research," *Frontiers in Psychology*, vol. Volume 11 - 2020, 2020.

[14] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, and L.-P. Morency, "Simsensei kiosk: a virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS '14, (Richland, SC), p. 1061–1068, International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[15] F. S. Nahm, "Receiver operating characteristic curve: overview and practical use for clinicians," *Korean Journal of Anesthesiology*, vol. 75, pp. 25–36, Feb 2022.