Dash-M5H: An Interactive Dashboard for Multi-Modal, Multi-Model Mental Health Assessment

Raymond Alavo[1] **Xinyuan Zhang**[1] **Gemza Ademaj**[1] **Junhui Cai**[1]
Hyeokhyen Kwon [2] **Robert Cotes**[3] **Gari Clifford**[2] **Ahmed Abbasi**[*1]
[1]Human-centered Analytics Lab, University of Notre Dame
[2]Department of Biomedical Informatics, School of Medicine, Emory University
[3]Department of Psychiatry and Behavioral Sciences, School of Medicine, Emory University
aabbasi@nd.edu
https://dash-m5h.io

Abstract

We present **Dash-M5H**, an interactive dashboard for *multi-modal, multi-model mental health* assessment that helps clinicians and researchers jointly inspect multimodal behavioral data with multi-model signal outputs of recorded clinical interviews. Guided by signal detection and integrated sensemaking theories, Dash-M5H synchronizes transcript text, audio, and facial behavior (action units and gaze) to support overview-to-detail evidence tracing; and it integrates extracted signals (e.g., sentiment and facial activity) with a clinically grounded VLM prediction pipeline that produces DSM-5-aligned depression predictions. Dash-M5H[1] is implemented in a lightweight, browser-based stack (`Quarto + Observable JS + D3`), supports local data import and time-synced clinical annotation with export. We demonstrate Dash-M5H through a depression screening scenario, evaluate its note-taking and screening capabilities through a user experiment, and release a live demo[2] and code[3] to facilitate reproducible evaluation.

1 Introduction

According to the World Health Organization, approximately one billion people worldwide live with a mental health disorder. Depression (280 million people) and anxiety (301 million) are the most prevalent conditions (Cuijpers et al., 2023). However, the availability of psychiatric care is limited even in high-income countries (Seyedi et al., 2023). This global challenge has drawn growing attention, with digital systems aiming to support both patients and clinicians. Prior work highlights patient-facing tools for self-screening and self-help as complements to professional care (Balcombe and De Leo, 2022), proposes richer diagnostic approaches that capture lived experience beyond standard questionnaires (Veldmeijer et al., 2023), and reviews clinician-facing decision support systems (Tong et al., 2025). These lines of work motivate interactive tools that connect patient experience, interpretable summaries, and clinical workflow—an aim our dashboard is designed to support.

Mental health assessment is typically conducted through a structured conversation in which clinicians probe multiple life domains while attending to cues in language, vocal prosody, and nonverbal behavior (Cotes et al., 2022). As recording becomes more common, sessions can be revisited for follow-up and analysis. However, review remains labor-intensive: clinicians must scrub through recordings, cross-reference transcripts, and synthesize information that occurs across modalities and timescales (Seyedi et al., 2023). Meanwhile, machine learning models can extract behavioral signals from raw data, such as facial action units and sentiment trajectories, and infer depression from multi-modal inputs (Jiang et al., 2024b; Qin et al., 2025a). Designing effective support therefore requires tools that facilitate *signal detection* from multi-modal data and support *integrated sensemaking* by bringing extracted signals and predictive outputs into a single view that clinicians can inspect and interpret jointly.

To address this gap, we introduce Dash-M5H, an integrated multi-modal, multi-model dashboard for mental health assessment that supports session-level overview and minute-level inspection by aligning transcripts, audio, and facial behavior signals with model-derived cues and predictions in a single interface. Dash-M5H is built for rapid session review and supports clinician note-taking and diagnosis. Our key contributions include:

- **An integrated multi-modal multi-model system for mental-health assessment** that synchronizes transcript, audio, and facial behav-

[*]Corresponding author.
[1]https://dash-m5h.io
[2]https://youtu.be/w3qCJ02k6bw
[3]https://github.com/nd-hal/M5H-Dashboard-VLM

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 138–147

ior signals with multi-model outputs, supporting overview-to-detail evidence tracing from session-level patterns to minute-level inspection.
- **Clinically grounded VLM prediction pipeline** that produces DSM-5-aligned depression predictions via a fine-tuned vision-language model for high-accuracy decision support.
- **A lightweight, reproducible, browser-based implementation** that supports local data import, structured time-stamped notes, optional interaction logging, and note export.

The remainder of this paper describes related work (§2), Dash-M5H interface (§3), system overview (§4), system architecture and implementation (§5), and a case study and user experiment (§6).

2 Related Work

Prior work approaches the mental health support challenge from complementary angles, spanning patient-facing tools and clinician-facing diagnostic decision support. On the patient side, Balcombe and De Leo (2022) discuss digital tools for self-screening, early detection, and self-help. Examples include mental health chatbots (Gabriel et al., 2024; Qiu et al., 2024). From the clinician and diagnostic perspective, recent work has focused on capturing and predicting patient mental health status based on remote interview transcripts (Cotes et al., 2022; Seyedi et al., 2023; Jiang et al., 2024b,a; Qin et al., 2025a). Veldmeijer et al. (2023) argue that while questionnaires and interviews are useful for symptom identification, further tools are needed to represent the complexity and idiosyncrasy of an individual's experience; namely, clinical decision support systems (CDSS) that offer richer context. Tong et al. (2025) review mental-health CDSS through the lens of healthcare professionals, noting recurring limitations including lack of transparency, interpretability, and granularity that undermine trust and inhibit adoption. These findings suggest an opportunity for systems that combine patient-centered information gathering with clinician-facing, interpretable representations and actionable views that support note-taking and diagnosis. Dash-M5H contributes to this need by providing time-synchronized, coordinated views of language, voice, and facial cues that help clinicians quickly locate moments and drill down into supporting evidence.

3 The Dash-M5H Interface

Figure 1 shows the user interface of the proposed Dash-M5H dashboard.

3.1 Global navigation

Dash-M5H includes a toolbar at the top with a global time control. A minute-level slider drives coordinated updates across the transcript, audio, and detailed facial views. When the user navigates to a given minute, the transcript highlights the corresponding speech segment, the audio player jumps to the relevant timestamp, and the single-face heatmap updates to reflect facial activations at that time. Overview visualizations (e.g., the Face Strip) remain visible to preserve whole-session context. When users expand a card, the interface expands the visualization to reveal more detail.

3.2 Row 1: Text and sentiment

Row 1 supports transcript-based review. Stacked sentiment bars summarize the frequency (or proportion) of emotion labels over time. Each bar includes a small highlight strip at the bottom indicating sentiment model confidence. Hovering over a bar shows the corresponding emotion frequency and confidence score, and clicking a bar updates the global time to that minute.

The transcript panel provides a time-synced text view in which emotion-related words are color-coded using the same emotion labels. Together, these views help users find moments when the text and inferred emotion change, serving as a starting point for inspecting the audio and facial signals.

3.3 Row 2: Facial expressions

Row 2 provides facial behavior evidence at two levels of detail. The Face Strip is a minute-level grid of facial expression, each summarizing the dominant facial activation for a minute of the session. The currently active minute is highlighted, and hovering reveals confidence scores from the underlying facial analysis model.

Selecting a minute updates the detailed single-face view, which overlays a muscle activation heatmap on a neutral face wireframe. This setup supports an overview-to-detail workflow: users can scan for key moments first, and then zoom in to examine the underlying action units.

3.4 Row 3: VLM predictions and audio

Row 3 shows VLM predictions and vocal evidence. The prediction panel displays a DSM-5-aligned

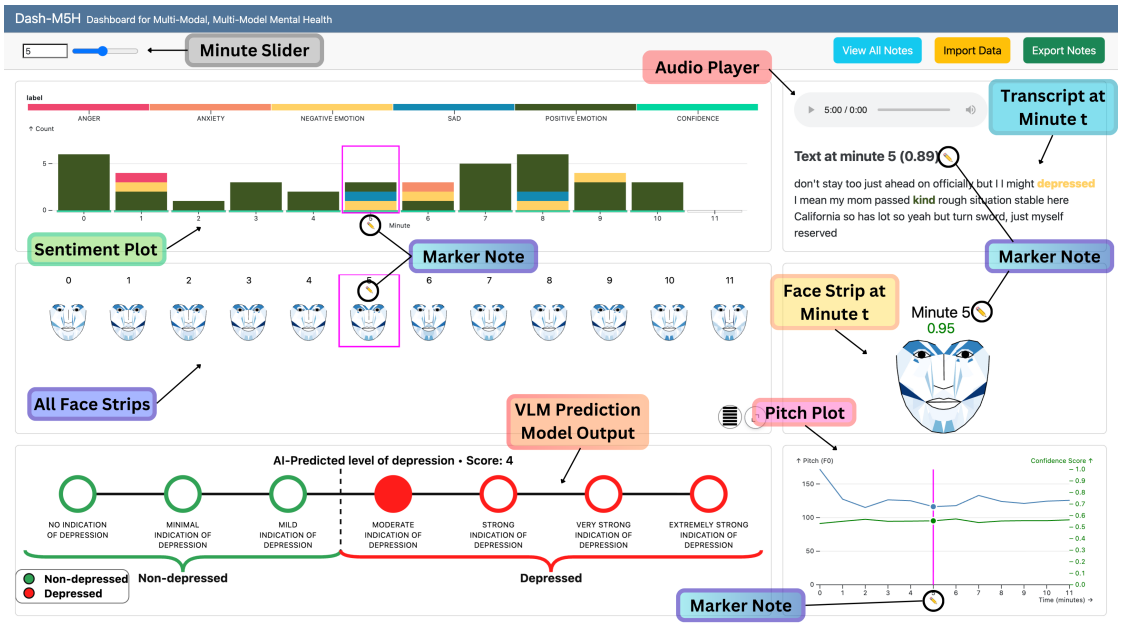


Figure 1: Screenshot overview of the Dash-M5H dashboard. *The dashboard is optimized for non-mobile browsers.*

depression level prediction on a Likert-style scale (0–7), produced by a fine-tuned vision–language model (VLM) for high-accuracy decision support (see Appendix A). The prosody view visualizes pitch over time and is synchronized with an embedded HTML5 audio player. When the global time control changes, the player jumps to the corresponding timestamp, enabling users to listen to the selected segment while inspecting pitch dynamics. These linked views help users inspect vocal markers like flat intonation or abrupt pitch changes.

3.5 Clinical annotation and export

To support clinician workflows, Dash-M5H includes an annotation system that ties note-taking to time and modality. Users can create notes directly from each panel, and note markers are rendered on the corresponding panels to indicate when and where the notes were taken. Notes can be edited or deleted in a history panel, and "View All Notes" panel aggregates all notes in chronological order. Finally, the complete set of session notes can be exported as a CSV file for documentation, supervision, or research.

4 System Overview

Figure 2 illustrates the system design of Dash-M5H, and Table 1 summarizes the major components of the its pipeline.

Component	Description
Kernel theory (design rationale)	Signal Detection Theory (access to raw social/technical cues) + Integrated Sensemaking Theory (identify systematic fluctuations and consistent characteristics); motivates cross-modality comparison, access to VLM predictions, and discrete granularity levels.
Inputs	Socio cues: video and audio; facial recognition, speech-to-text, pitch analysis.
Pre-processing/ Alignment	Offline ML pipeline: LIWC sentiment analysis, OpenFace facial behavior extraction, pitch analysis; standardize session window and time-align signals; compute summaries (sentiment bars, facial heatmap, pitch fluctuation).
VLM prediction	Fine-tuned vision-language model for DSM-5-aligned depression level prediction.
Interactive dashboard	Time-stamped transcript; sentiment view; audio player; pitch fluctuation; facial overview and focal heatmaps; VLM prediction score.
User actions	Navigate by time (scrub/jump); select a minute to update focal views; hover/click to inspect details; play audio at selected time; write/organize clinical notes during review.
Outputs	Clinical notes organized by assessment dimensions (e.g., appearance, speech, emotion); depression assessment support via VLM prediction; exportable notes/report (e.g., CSV).

Table 1: Summary of Dash-M5H pipeline.

4.1 Kernel theories

Dash-M5H is designed for mental health professionals (and researchers) who review recorded assessment sessions and need to integrate evidence across language, voice, and facial behavior (Ademaj et al., 2025). Figure 2 summarizes Dash-M5H as a theory-grounded pipeline (Lalor et al., 2025) that links multi-modal assessment signals to coordinated views and structured documentation. At the top, the design is informed by two kernel the-

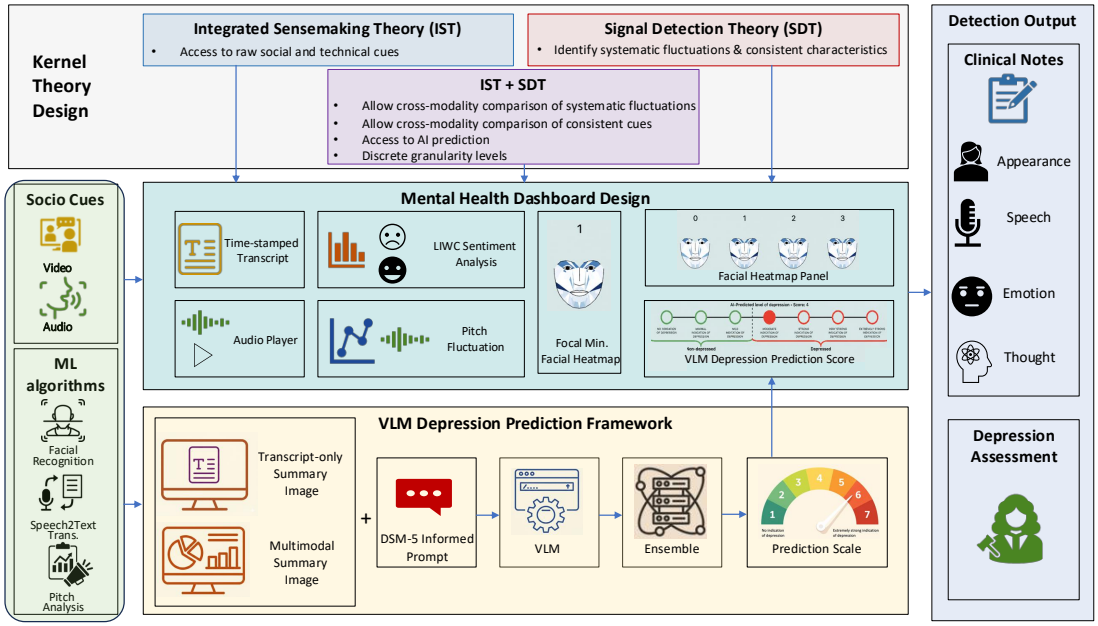


Figure 2: System design of the Dash-M5H dashboard.

ories: (i) *Integrated Sensemaking Theory*, which emphasizes access to raw social and technical cues (Maitlis et al., 2013), and (ii) *Signal Detection Theory*, which emphasizes interpreting systematic fluctuations and consistent characteristics over time (Lynn and Barrett, 2014). This framing motivates three interface requirements: cross-modality comparison, overview-to-detail patterns tracing, and access to accurate prediction.

4.2 Inputs and upstream processing

On the left of Figure 2, Dash-M5H ingests socio-cues (video and audio) alongside pre-computed ML-derived signals. Concretely, the dashboard expects modality-specific CSV files such as a time-stamped transcript, audio prosody features (e.g., pitch), and OpenFace-derived facial signals (e.g., action units and gaze). To make modalities comparable, Dash-M5H standardizes sessions to a fixed window (e.g., 12 minutes) and aligns signals to a shared timeline (e.g., 1-minute bins, optionally finer), producing derived summaries such as sentiment bars, pitch traces, and AU activity summaries.

4.3 Integrated views and interactions

The center of Figure 2 depicts the Dash-M5H dashboard as a set of coordinated, time-synchronized views controlled by a single global time control. This global control drives "details-on-demand" views, enabling rapid navigation while preserving whole-session context. Clinicians can quickly scan for salient moments (e.g., shifts in inferred affect, abrupt pitch changes, or bursts of facial activity) and then drill down using detailed views, synchronized audio playback, and focal facial visualizations. The dashboard also integrates VLM depression predictions. Throughout review, users can add time-synced notes directly in the interface and navigate between modalities using linked interactions.

4.4 Clinically grounded VLM prediction framework

The bottom portion illustrates the VLM-based depression prediction framework for predicting DSM-5-aligned depression level. The framework takes two session summary images as input: a transcript-only summary image and a multi-modal summary image that aggregates extracted behavioral cues. We then apply a DSM-5-informed prompt to guide the model's reasoning and fine tune the model. The VLM outputs a depression level score on a 1–7 scale. VLM-based scoring was employed because it was found to outperform fine-tuned BERT and LLama-3-70b models (See Appendix A).

4.5 Clinical outputs

On the right, Figure 2 emphasizes documentation-oriented outputs. The system supports structured clinical notes (e.g., aligned with categories such as appearance, speech, affect/emotion, and thought content) and consolidates notes into one session report for export (CSV). Users can annotate findings directly within the original environment and

export a session report. The dashboard is intended to support review and documentation workflows and does not replace clinical judgment.

5 System Architecture and Implementation

Building on the system design, Figure 3 shows the architecture and data flow of Dash-M5H.

M5H was developed on a lightweight, browser-based architecture intended to be very responsive. The interface is implemented in Quarto, which combines Markdown-based narrative with embedded `Python/R` for preprocessing and Observable JavaScript for reactive client-side components. Custom `D3.js` visualizations render high-fidelity `SVG` views for the transcript, audio curves, and facial displays.

The dashboard is organized into modular JavaScript components: (i) Face plotting component to render the Face Strip and the detailed single-face heatmap from OpenFace coordinates; (ii) Note component that manages time-synced annotations, history, CRUD operations, and CSV export; and (iii) Tracking component that optionally logs user interactions (mouse coordinates and click events) to a telemetry endpoint for research on inspection behavior.

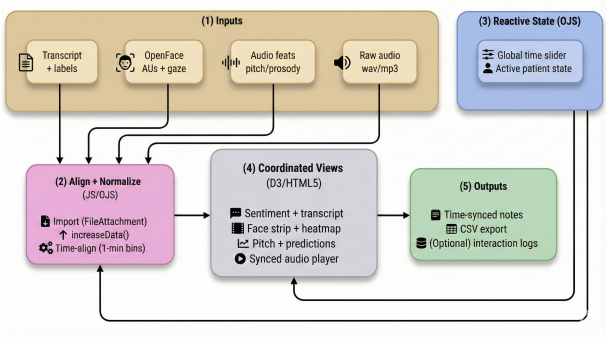


Figure 3: Data flow and system architecture of the Dash-M5H dashboard.

5.1 Data ingestion, import, and normalization

By default, Dash-M5H loads pre-processed `CSV` files for transcript, facial signals, gaze, and audio features using the Observable API, which makes it straightforward to swap datasets. For broader usability, the dashboard also provides an "Import Data" workflow that allows users to override the default files by uploading local `CSV`s and an audio recording (e.g., `.mp3/.wav/.ogg`) as depicted by Figure 3 in the Input block. Uploaded files are read client-side and injected into the reactive state, enabling updates without page refresh.

To ensure consistent visualization across sessions, Dash-M5H normalizes each interview to a fixed 12-minute window. Longer sessions are truncated and shorter sessions are padded (e.g., with neutral-face placeholders and empty bins) so that overview visualizations such as the Face Strip remain comparable across the cohort. This happens at step 2 of our flow diagram as shown in Figure 3.

All the views are coordinated by a reactive slider that orchestrate the dashboard views. Annotation can be made at every timestamp of each view. They can be viewed, edited, and deleted before being exported.

5.2 Optional interaction telemetry

Dash-M5H can optionally collect interaction telemetry to support usability studies and to better understand how clinicians inspect multimodal evidence. The tracking module records viewport-relative mouse coordinates at 1 Hz and click events, and forwards them to an AWS API Gateway endpoint. Viewport mapping ensures that interaction traces remain comparable across devices and when the dashboard is embedded (e.g., in an iframe). For privacy, telemetry can be disabled at deployment time, and the demo can be configured to run entirely locally without transmitting any interaction data.

6 Evaluation of Dash-M5H

6.1 Case study: Depression assessment

We present a case study using a patient with depression in Figure 4 to illustrate the dashboard's integrative sensemaking capabilities. After importing the multimodal data and selecting the patient, the dashboard synchronizes all visualizations at the minute level. In this case, the VLM depression prediction displays a score of 6 (highlighted in red), suggesting a very strong indication of depression. This prediction serves as a reference point while the dashboard supports joint inspection of multi-modal multi-model patterns across sentiment, transcript, facial activation, and pitch.

Multimodal behavioral visualizations: To examine whether the VLM-based prediction of strong depression is supported by the behavioral evidence, the user first scans the sentiment bar charts to identify recurring affective patterns. During this overview, negative categories increase early in the

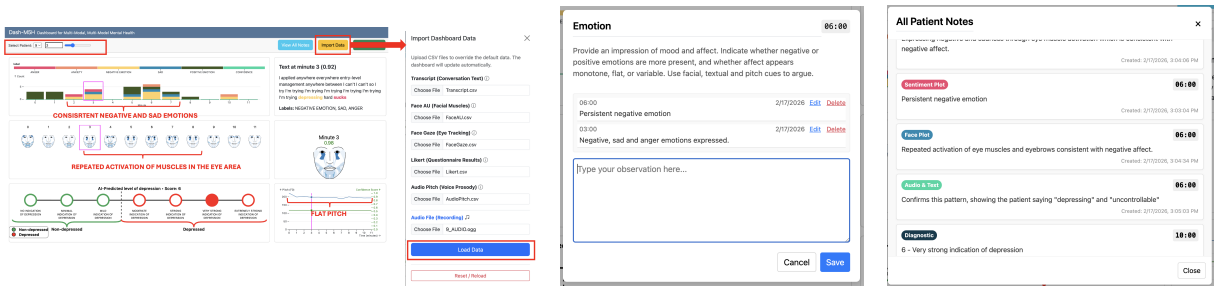


Figure 4: Example use cases of Dash-M5H. **Left**: Importing session data and reviewing multimodal behavioral visualizations. **Middle**: Structured, modality-specific note-taking during review. **Right**: Reviewing, consolidating, and exporting documentation.

interview. Around minute 3, the stacked bar chart shows a noticeable co-occurrence of negative emotion, sadness, and anger. To investigate this segment more closely, the user navigates to minute 3. The selected minute is highlighted in purple across all panels, aligning the transcript, facial activation, and pitch visualizations to the same time point. The transcript is synchronized to display the corresponding speech ("depressing," "sucks"), highlighting words labeled NEGATIVE EMOTION, SAD, and ANGER using the same color scheme as the sentiment bars. The facial activation map shows stronger activation in the eyebrow and eye regions, consistent with the negative affect indicated in the sentiment bars. The pitch visualization, marked at minute 3, indicates reduced variability. By aligning these signals temporally, Dash-M5H allows for cross-modality comparison and assessment of signal consistency. At this point, the multimodal evidence aligns with the VLM prediction.

Structured modality-specific note-taking: As these patterns are identified, Dash-M5H allows users to document observations within structured categories (Appearance, Speech, Emotion, Thought Content). At minute 3, the user documents observations across modalities: negative language from the transcript is recorded in the Emotion panel as "Negative, sad and anger emotions expressed," while the corresponding eye-region activation is documented in the Appearance/Behavior panel as "Expressing negative and sadness through eye muscle activation consistent with negative affect." Each note is time-stamped and tied to the selected minute and panel. By documenting observations as they occur, the user is able to keep track of the identified patterns across modalities and time points.

As navigation continues, additional notes are added at later minutes to document recurring patterns, producing a time-ordered record of the sense-making process. For this patient, the dashboard reveals limited positive emotion and reduced facial variability across minutes. The sentiment plot shows repeated negative and sad categories aligned with the transcript which contains negative expressions, facial activation remains concentrated in the brow and eye regions, and pitch stays relatively flat. Given that all these signals and cues, the user records a very strong indication of depression, consistent with the VLM prediction.

Reviewing and exporting documentation: After documenting observations across the interview, the user clicks "View All Notes" to review the assessment. The notes are displayed chronologically and labeled by visualization panel, making it easy to see when and where observations were documented. Reviewing the notes shows that negative emotion and eye-region activation were documented at multiple time points (e.g., minutes 3 and 6), indicating consistency across modalities and over time. This integrated view allows the user to see how observations were documented across minutes and modalities, providing a structured overview of the assessment process. The user can then select "Export Notes" to download the documentation as a `CSV` file, ensuring that the recorded observations remain organized and traceable for clinical documentation or further analysis.

6.2 User depression detection and note-taking performance

We report results on a user study evaluation of Dash-M5H for assessing depression, relative to a baseline dashboard devoid of multi-modal, multi-model and VLM outputs (see Figure 6). The study employed 55 healthcare professionals (n = 55) in Table 2. Of these, 22 participants were highly experienced in mental health screening (12 in Dash-M5H and 10 in baseline), while the remaining 33 had low prior experience (18 in Dash-M5H and 15 in baseline). Across different user expertise levels,

we evaluate how effectively Dash-M5H supports depression assessment compared to the baseline, with each participant assessing data for three patients. The participants using Dash-M5H show better performance on depression assessment. As an additional evaluation, we analyze participants' note-taking to assess their sensemaking during depression assessment (bottom of Table 2). We qualitatively coded notes as descriptive (based on a single modality or panel) or advanced, integrative (comparing or combining information across multiple modalities/panels), and report the percentage of notes in each category. Relative to the baseline, participants using Dash-M5H show more advanced integrative sensemaking, as their notes more often compare and combine signals across panels. This suggests that Dash-M5H supports more integrative interpretation during depression assessment, which is an important precursor for enhanced data-driven decision-making (Maitlis et al., 2013).

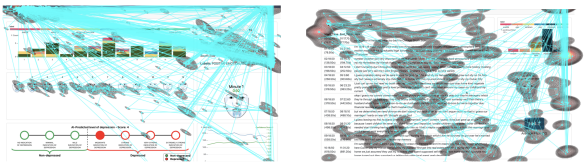
Table 2: Performance and sensemaking metrics across experience levels for Baseline and Dash-M5H.

	High Exp		Low Exp		Overall	
Metric	**Base**	**Dash**	**Base**	**Dash**	**Base**	**Dash**
Precision	43.8%	76.2%	68.4%	73.1%	57.1%	74.5%
Recall	58.3%	100.0%	61.9%	86.4%	60.0%	92.1%
F1	50.0%	86.5%	65.0%	79.2%	58.8%	82.4%
Desc. Sensemaking	48.3%	26.8%	47.1%	22.1%	47.9%	24.1%
Adv. Sensemaking	51.7%	73.2%	52.9%	77.9%	52.1%	75.9%

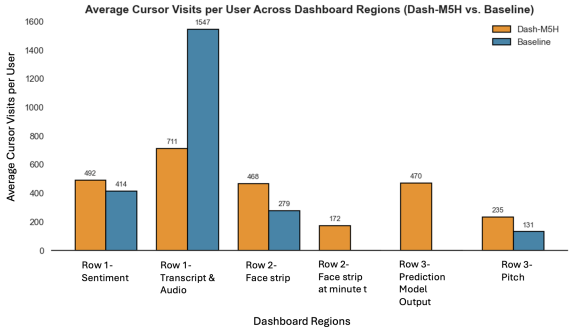
Precision, Recall, and F1 denote classification performance. Descriptive (Desc). and Advanced (Adv.) Integrative Sensemaking reflect types of sensemaking behaviors across participant groups based on percentage of notes in each category.

6.3 Assessment interface interaction

We show the cursor movement patterns from the user study ($n = 55$) to quantify where participants focus their attention during the assessment. Mouse tracking data indicates that in the Dash-M5H condition, cursor activities are more distributed across the different visualization panels, with frequent transitions between signals (Figure 5(a)). Figure 5(c) shows that participants using Dash-M5H interact with the audio player and transcript, as well as the sentiment and face strip plots and the prediction model outputs. In contrast, in the baseline condition, cursor movements are more concentrated on a limited set of interface elements, see Figure 5(b), mainly the audio player and transcript, see Figure 5(c). This suggests that Dash-M5H facilitates attending to a broader set of information sources during the assessment, contributing to advanced sense-making and enhanced diagnosis performance.



(a) Dash-M5H. (b) Baseline.



(c) Average cursor visits per user across dashboard regions (Dash-M5H vs. Baseline).

Figure 5: Mouse movements while assessing depression. Cursor visit values are normalized per user within each testbed and regions are aligned with the dashboard layout (Row 1–Row 3).

7 Conclusion and Future Work

We present Dash-M5H, an interactive browser-based dashboard for multi-modal, multi-model mental health assessment. The dashboard operates in a local-first mode where imported data, derived features, and annotations never leave the user's machine, preserving privacy. Future work will add real-time streaming, richer vision–language explanations, and longitudinal clinician user studies to quantify impacts on review quality and efficiency.

Acknowledgments

Ahmed Abbasi was partially funded by an Industry Labs Seed Grant. Dr. Kwon is partially funded by the National Institute of Health (1R21DC021029), Georgia CTSA, and Shriners Hospital for Children.

Ethics Statement

This work, including the user experiment, was reviewed and approved by the University of Notre Dame IRB (Protocol #25-03-9149). The VLM predictions are intended to serve as a decision support to healthcare professionals, and are **not** diagnostic, are **not** used to make clinical determinations, and should always be verified against the underlying audio/visual evidence using professional human judgment. When designing the predictive elements and dashboard components, we adhered to the tenets of responsible AI for healthcare contexts

(Lalor et al., 2022, 2024; Oketch et al., 2025; Krishnan et al., 2025). To preserve patient privacy, we avoid sharing identifiable patient data; all examples and upload demonstrations use synthetic and/or de-identified data with identifying metadata removed. Access to any study materials is restricted to authorized personnel under standard security controls.

References

Gemza Ademaj, Xinyuan Zhang, Junhui Cai, Ahmed Abbasi, Saonee Sarker, and Suprateek Sarker. 2025. Designing support for sensemaking in multimodal, multi-model mental health assessments. In *Forty-Sixth International Conference on Information Systems (CIS)*, 15, pages 1–9.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Luke Balcombe and Diego De Leo. 2022. Human-computer interaction in digital mental health. *Informatics*, 9(1).

Robert O Cotes, Mina Boazak, Emily Griner, Zifan Jiang, Bona Kim, Salman Seyedi, Ali Bahrami Rad, and Gari D Clifford. 2022. Multimodal assessment of depression utilizing video, acoustic, and heart technology: Protocol for an observational study. *JMIR Research Protocols*, 11(7):e36417.

Pim Cuijpers, Afzal Javed, and Kamaldeep Bhui. 2023. The who world mental health report: a call for action. *The British Journal of Psychiatry*, 222(6):227–229.

Saadia Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. 2024. Can ai relate: Testing large language model response for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2206–2221.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, and 1 others. 2014. The distress analysis interview corpus of human and computer interviews. In *Lrec*, volume 14, pages 3123–3128. Reykjavik.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Zifan Jiang, Salman Seyedi, Emily Griner, Ahmed Abbasi, Ali Bahrami Rad, Hyeokhyen Kwon, Robert O Cotes, and Gari D Clifford. 2024a. Evaluating and mitigating unfairness in multimodal remote mental health assessments. *PLOS Digital Health*, 3(7):e0000413.

Zifan Jiang, Salman Seyedi, Emily Griner, Ahmed Abbasi, Ali Bahrami Rad, Hyeokhyen Kwon, Robert O Cotes, and Gari D Clifford. 2024b. Multimodal mental health digital biomarker analysis from remote interviews using facial, vocal, linguistic, and cardiovascular patterns. *IEEE journal of biomedical and health informatics*, 28(3):1680–1691.

Ramayya Krishnan, John P Lalor, Nicolas Prat, and Ahmed Abbasi. 2025. From policy to practice: Research directions for trustworthy and responsible ai 'by design'. *IEEE Intell. Syst.*, 40(5):45–51.

John P Lalor, Ahmed Abbasi, Kezia Oketch, Yi Yang, and Nicole Forsgren. 2024. Should fairness be a metric or a model? a model-based framework for assessing bias in machine learning pipelines. *ACM Transactions on Information Systems*, 42(4):1–41.

John P Lalor, Ruiyang Qin, David Dobolyi, and Ahmed Abbasi. 2025. Textagon: Boosting language models with theory-guided parallel representations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 82–92.

John P Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in nlp. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 3598–3609.

Yanhong Li, Zixuan Lan, and Jiawei Zhou. 2025. Text or pixels? it takes half: On the token efficiency of visual text inputs in multimodal llms. *arXiv preprint arXiv:2510.18279*.

Spencer K Lynn and Lisa Feldman Barrett. 2014. "Utilizing" signal detection theory. *Psychological science*, 25(9):1663–1673.

Sally Maitlis, Timothy J Vogus, and Thomas B Lawrence. 2013. Sensemaking and emotion in organizations. *Organizational Psychology Review*, 3(3):222–247.

Kezia Oketch, John P Lalor, Yi Yang, and Ahmed Abbasi. 2025. Bridging the llm accessibility divide? performance, fairness, and cost of closed versus open llms for automated essay scoring. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 655–669.

Ruiyang Qin, Kai Yang, Ryan Cook, Salman Seyedi, Emily Griner, Hyeokhyen Kwon, Robert Cotes, Zifan Jiang, Gari Clifford, and Ahmed Abbasi. 2025a. Language models for online depression detection: A review and benchmark analysis on remote interviews. *ACM Transactions on MIS*, 16(2):1–35.

Wei Qin, Zetong Chen, Xun Yang, Lei Wang, Yunshi Lan, Weijieying Ren, and Richang Hong. 2025b. Explainable and interactive llms-augmented depression detection in social media. *IEEE Transactions on Computational Social Systems*.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636.

Salman Seyedi, Emily Griner, Lisette Corbin, Zifan Jiang, Kailey Roberts, Luca Iacobelli, Aaron Milloy, Mina Boazak, Ali Bahrami Rad, Ahmed Abbasi, and 1 others. 2023. Using hipaa compliant transcription services for virtual psychiatric interviews: Pilot comparison study. *JMIR Mental Health*, 10:e48517.

Fangziyun Tong, Reeva Lederman, and Simon D'Alfonso. 2025. Clinical decision support systems in mental health: A scoping review of health professionals' experiences. *International Journal of Medical Informatics*, 199:105881.

Lars Veldmeijer, Gijs Terlouw, Job van 't Veer, Jim van Os, and Nynke Boonstra. 2023. Design for mental health: can design promote human-centred diagnostics? *Design for Health*, 7(1):5–23.

A VLM Prediction Framework

Dash-M5H includes a prediction pipeline designed for the multimodal mental health data under the multi-model environment, with the goal of providing the dashboard AI depression prediction results that are highly accurate and best support user sensemaking process. We choose Qwen2.5-VL-72B, an open-sourced large vision-language model (VLM) that can understand text, images, and video (Bai et al., 2025). We provided multimodal information for each session with a session summary image as shown in Figure 6, where the left side displays timestamped full transcript, and the right side provides a bar chart of emotional words frequency count, the average facial expression, and average pitch. We designed the prompt to include the general definition of depression, an abbreviated version of the DSM-5 criteria (Qin et al., 2025b), assessment guidelines, and a 1-7 Likert rating scale. The model is asked to generate a depression indication scale between 1 and 7, and a short explanation of their assessment reasoning. The example prompt can found in Appendix A.

Moreover, while interview transcript naturally contains most information comparing to other modalities, we also leverage VLM's ability to process textual content particularly. Research has shown that rendering long text as an image and feeding it to a MMLM reduces decoder tokens by approximately half while maintaining task performance (Li et al., 2025). Hence, we provide a second set of data in the form of transcript-only images, where each image contains only the left panel in Figure 6. The prompt was designed similarly with the only difference of including a sentence describing different key components in the image.

We fine tune and evaluate the models using 5-fold cross-validation on DAIC-WOZ and Extended DAIC-WOZ testbeds (Gratch et al., 2014). The Low-Rank Adaption (LoRA) technique (Hu et al., 2022), a parameter-efficient method that freezes the pre-trained model weights and injects trainable low-rank decomposition matrices into specific layers of the transformer architecture. For both models, we applied LoRA adapters to the query, key, value, and output projection matrices within the attention mechanism as well as the gate, up, and down projection matrices in the feed-forward network. We integrate the results from both procedures and obtain the final prediction through a balanced ensembling approach. We achieve accuracy of 0.867,

macro F1 of 0.866, and AUC of 0.902. Table 3 presents a summary of prediction results.

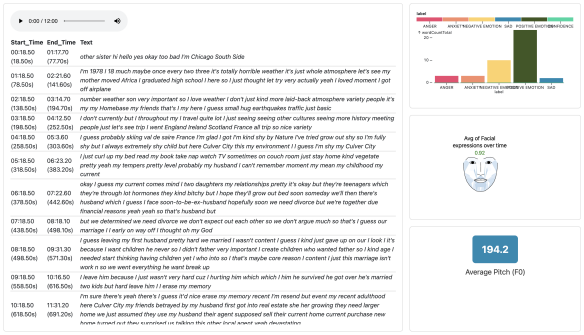


Figure 6: Baseline dashboard.

Method	AUC	Accuracy	Macro F1
Ensemble (Qwen2.5-VL 72B)	**0.902**	**0.867**	**0.866**
Qwen2.5-VL 72B (zero-shot)	0.844	0.783	0.783
Llama3 70B (fine-tuned)	0.852	0.833	0.833
BERT (fine-tuned)	0.548	0.500	0.438

Table 3: Summary results for AI depression prediction.

Prompt for Depression Assessment

Task: You are conducting a mental health assessment for Patient P using clinical guidelines.
CRITICAL ASSESSMENT GUIDELINES:

- Base your assessment ONLY on what you can clearly observe in the image
- Do NOT invent, assume, or extrapolate details not clearly visible
- If text is unclear or illegible, note this rather than guessing content
- Use DSM-5 criteria below as guidance for identifying concerning patterns

You are analyzing ONE mental health dashboard image from a 12-minute clinical session. IMAGE (Session Summary): Contains the full conversation transcript, average facial expression analysis where facial action unit movements are highlighted, confidence scores of how accurate the data captured are, and average pitch measurements. NOTE for IMAGE: The score above the average facial expression plot is NOT the expression score; it is the score of how accurate facial movements are captured. The score DOES NOT have direct indications of emotions. DSM-5 MAJOR DEPRESSIVE DISORDER CRITERIA (for reference): Look for evidence of these patterns in the observable data:

1. Depressed mood (sadness, emptiness, hopelessness in transcript/expressions)
2. Loss of interest/pleasure in activities (flat affect, minimal engagement)
3. Appetite changes (may not be observable in this data)
4. Sleep disturbances (may not be observable in this data)
5. Psychomotor agitation or retardation (observable in expressions/speech patterns)
6. Fatigue or loss of energy (low pitch, flat expressions, slow speech)
7. Feelings of worthlessness or guilt (observable in transcript content)
8. Decreased concentration (observable in speech patterns, conversation flow)
9. Thoughts of death/suicide (observable in transcript content)

RATING SCALE:

1 No indication of depression based on DSM-5 criteria
2 Minimal indication of depression based on DSM-5 criteria
3 Mild indication of depression based on DSM-5 criteria
4 Moderate indication of depression based on DSM-5 criteria
5 Strong indication of depression based on DSM-5 criteria
6 Very strong indication of depression based on DSM-5 criteria
7 Extremely strong indication of depression based on DSM-5 criteria

YOUR RESPONSE MUST USE THIS EXACT FORMAT:
Depression Indication Scale: [NUMBER 1–7]
Clinical Reasoning: 2–3 sentences explaining your rating based ONLY on clearly observable patterns in the image. Reference specific DSM-5 criteria if present, or explain why patterns fall within normal range. Be specific about what you can actually see versus what might be unclear.