Summary of the BeyondSmile Challenge on Detecting Depression Through Facial Behavior and Head Gestures

Rahul Islam* [1], Tongze Zhang [2], Anlan Dong [3], Melik Ozolcer [4], Christina Garcia [5], Milyun Ni'ma Shoumi [6], Prerna Jamloki [7], Sozo Inoue [8], Sang Won Bae [9]

[1]Emory University Atlanta, USA

[2349]Stevens institute of Technology Hoboken, USA

[5678] Kyushu Institute of Technology, Kitakyushu, Japan

Abstract

Early detection of depressive episodes is crucial in managing mental health disorders such as Major Depressive Disorder (MDD) and Bipolar Disorder. However, existing methods often necessitate active participation or are confined to clinical settings. While facial expressions have shown promise in laboratory settings for identifying depression, their potential in real-world applications remains largely unexplored. In this challenge we introduce a data collected state of the art facial behavior sensing system, that tracks different facial behavior primitives such as Action Units, Landmarks, Head Pose, Eye Open state and others for task of detecting depressive episode in the wild. The challenge duration was three months, from Dec 12, 2024 to Feb 28, 2025 and 8 teams participated. Among the final teams, Team Persistence (Team ID 160) achieved 0.77 AUROC for universal model and 0.88 AUROC for hybrid model and became the winner.

[1]rahul.islam@dbmi.emory.edu
[2]tzhang64@stevens.edu
[3]adong@stevens.edu
[4]mozolcer@stevens.edu
[5]alvarez7.christina@gmail.com
[6]milyun.nima.shoumi@gmail.com
[7]jamloki-prerna951@mail.kyutech.jp
[8]sozo@brain.kyutech.ac.jp
[9]sbae4@stevens.edu

1 Introduction

In recent years, there has been growing research in building low-burden passive sensing systems for mental health monitoring. A particular interest has been gained in monitoring depression through passive sensors in smartphones like GPS, screen, and others [2, 3, 5, 8, 15, 18, 22, 24, 27, 31]. COVID-19 driven social-distancing measures accelerated the use of telehealth, including telepsychiatry [6, 23]. Although anxiety and depression peaked across all age groups in 2020 and began easing in early 2021 [32], roughly 12 million U.S. adults (about 5 percent of the population) still cope with both chronic pain and clinically significant mood symptoms [17]. These realities highlight the need to rethink timely mental-health delivery; scalable, personalized psychiatry that emphasizes prevention and tailored interventions remains a key, but still scarce, solution [1].

Growing evidence shows that depression alters non-verbal cues, specifically facial muscle activity and head movements, or "facial-behavior primitives" [7, 9, 10, 29]. These involuntary markers enable automated systems to deliver objective, repeatable assessments of depressive symptoms [5, 28, 30]. Yet most studies rely on controlled lab videos, and real-world uptake is still hampered by privacy concerns, high costs, and significant computational demands [2, 3, 5, 8, 18, 22, 24, 27, 31].

Facial actions are well-established laboratory markers of depression, yet their deployment "in the wild" has been limited by the lack of lightweight, privacy-preserving mobile tools. We therefore in this challenge introduce a dataset collected from system FacePsy: an open-source smartphone framework [14] that passively extracts facial landmarks, eye-openness, smiles, and head-movement cues in real time while safeguarding user privacy. We posit that these digital facial biomarkers can reveal users' affective states and, in turn, enable on-device algorithms to flag depressive episodes. A naturalistic field study with FacePsy tests this premise by evaluating how well these signals predict depression outside controlled settings.

In addition, this work is presented within the context of the ABC 2025 Conference Coding Challenge [4, 12, 34], which aims to advance activity and behavior analysis through diverse real-world and synthetic datasets. By contributing a mobile, privacy-preserving facial-behavior dataset derived from FacePsy, this challenge track expands the scope of Human Activity Recognition (HAR) research toward more realistic, ecologically valid, and clinically meaningful applications. These initiatives collectively accelerate progress toward scalable digital mental-health assessment tools.

2 Challenge Design and Dataset

The BeyondSmile Challenge was structured to assess the capabilities of modern computational models in detecting depressive episodes from naturalistic behavioral data. The design centered on a rich dataset and two distinct, rigorously defined evaluation protocols.

2.1 Dataset and Ground Truth

The challenge utilized the FacePsy dataset [14], an open-source collection of facial behavior and head gesture data captured "in the wild." The data was collected using an Android application that opportunistically initiates a 10-second recording session when a user performs trigger actions, such as unlocking their phone. Over the course of the study, approximately 16,000 facial-capture events were recorded from 25 participants. After filtering for self-reported non-use periods and data quality issues, a total of 544 usable participant-days remained for analysis.

The ground truth for labeling depressive episodes was derived from the 9-item Patient Health Questionnaire (PHQ-9), a standard clinical screener for depression. As detailed in the challenge guidelines, a participant's two-week observation period is labeled as a depressive episode (label=1) if and only if their PHQ-9 score was ≥ 5 at both the start and end of the period; otherwise, it is labeled as non-depressive (label=0). After data cleaning and removing non-compliant cases, this process yielded 44 labeled intervals (14 depressive and 30 non-depressive) across 25 participants, forming the basis for the classification task.

2.2 Feature Primitives

Participants were provided with a rich set of low-level feature primitives extracted from each 10-second video recording session. These raw features, detailed in the challenge tutorial, served as the primary input for all subsequent feature engineering and modeling. The key primitives, extracted for each frame, included:

- Facial Action Units (AUs): A set of 12 AU intensities (e.g., AU06: Cheek Raiser, AU12: Lip Corner Puller) representing the activation of specific facial muscles, which are fundamental components of facial expressions.
- Classification Probabilities: Three real-time classification scores: `leftEyeOpenProbability`, `rightEyeOpenProbability`, and `smilingProbability`, providing direct, high-level indicators of facial state.
- Head Pose: Three head orientation metrics in the form of Euler angles (pitch, yaw, and roll), quantifying the movement and posture of the head in 3D space.
- Facial Landmarks and Contours: A dense set of 133 facial landmark coordinates (x, y) identifying key points on the face (e.g., corners of the eyes, nose tip, mouth outline) and the contours of facial components like the eyebrows and lips (see Figure 1).
- Bounding Box: The coordinates defining the location of the detected face in each frame, primarily used for visualization and image cropping.

2.3 Evaluation Protocols

Participants were tasked with building models for two distinct evaluation schemes, with the Area Under the Receiver Operating Characteristic

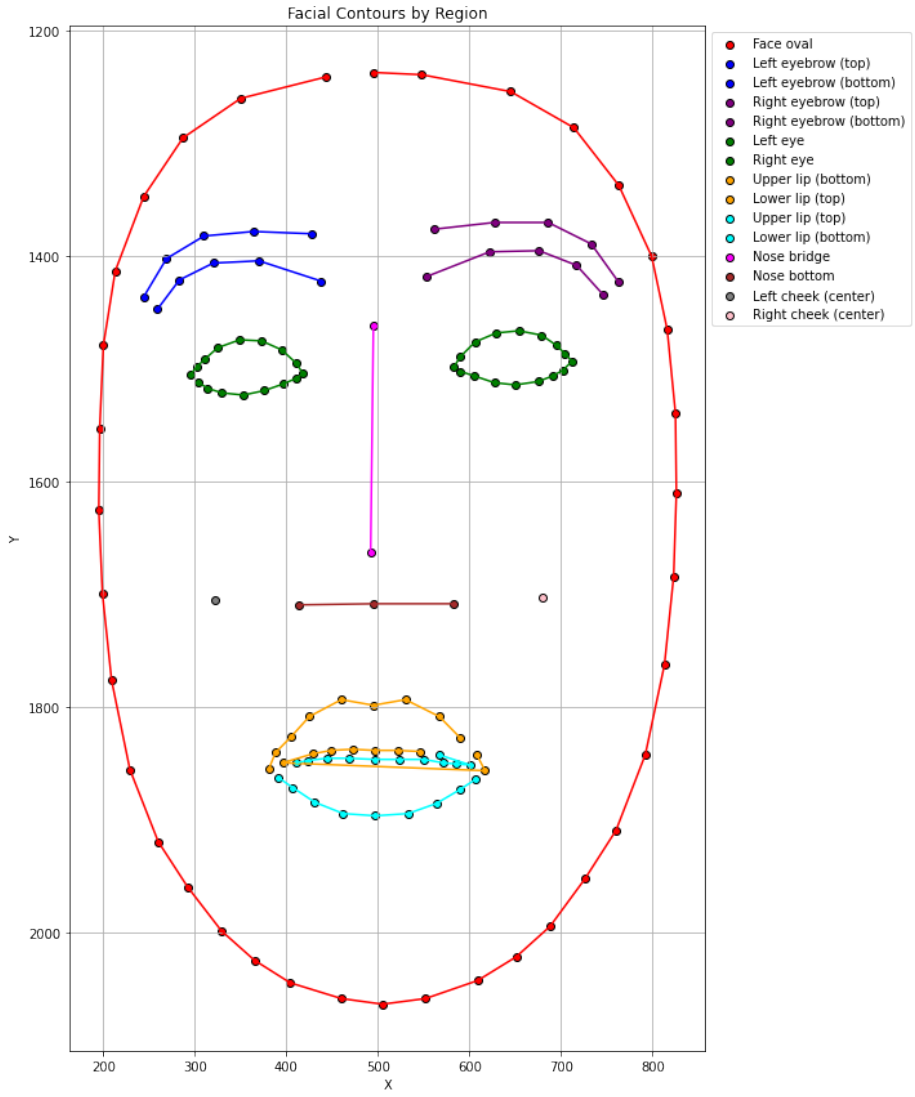

Figure 1: Visualization of Facial Feature Primitives. This figure illustrates some of the key low-level features provided to participants, including the 133 facial landmarks (blue dots), component contours (e.g., for eyes, eyebrows, and mouth, shown in different colors), and the face bounding box (not shown). These primitives form the basis for all subsequent feature engineering.

Curve (AUROC) as the primary performance metric.

1. Universal Model (LOPO): This track required teams to build a single, generalized model trained on data from all but one participant and tested on the held-out individual. This process was repeated for all participants using Leave-One-Participant-Out (LOPO) cross-validation. The goal was to assess a model's ability to generalize to completely new users, a critical requirement for any broadly deployable clinical tool.
2. Hybrid Model (LOPDO with Nested CV): This track aimed to simulate a real-world personalization scenario. It employed a time-series-aware Leave-One-Participant-Day-Out (LOPDO) nested cross-validation protocol. For a given test fold (e.g., data for participant P on day D), the model was trained on all data from other participants plus all available data from participant P on days prior to D. This protocol tests a model's ability to adapt and personalize its predictions by learning from a user's own history while leveraging a general population model.

3 Overview of Submitted Methodologies

The eight participating teams, representing a diverse international cohort from Indonesia, Vietnam, Japan, Poland, Pakistan, the UAE, and Bangladesh, submitted a wide array of solutions, which can be broadly categorized into three methodological themes: traditional machine learning pipelines, deep learning sequence modeling, and hybrid ensemble approaches. A summary of the key strategies is presented in Table 1.

3.1 Traditional Machine Learning Pipelines

A significant portion of teams relied on well-established machine learning pipelines that emphasized extensive, handcrafted feature engineering. AI-MED Squad (142) exemplified this approach by using time-series feature extraction libraries (tsfel, pycatch22) to generate a massive feature set from large 6-hour temporal windows. This strategy aimed to capture a wide range of statistical properties, which were then fed into classifiers like Random Forest and k-NN. In stark contrast, Lord of Depression (162) adopted a "minimalist" temporal strategy, extracting emotion and AU features from only the first frame of each 10-second recording, betting that a single, clean snapshot could be a powerful predictor.

3.2 Deep Learning and Sequence Modeling

Several teams explored deep learning architectures to automatically learn temporal representations from the data. Team Persistence (160) implemented Long Short-Term Memory (LSTM) networks, a natural choice for modeling sequential data like facial dynamics over time. Taking a more advanced approach, Team Access (119) and DP2K (154) utilized Transformer-based models. The solution from DP2K was particularly novel, as it constructed graphical representations of AUs and landmarks

within each frame and used a Transformer to explicitly model their complex inter-dependencies, followed by a second temporal Transformer to summarize the video. These approaches aimed to reduce the burden of manual feature engineering by learning relevant patterns directly from the data sequences.

3.3 Hybrid and Ensemble Methods

The third group of teams focused on combining the predictions of multiple models to enhance robustness and accuracy. Semicolon (161) implemented a classic stacked ensemble, using Extra Trees and XGBoost as base-level models whose predictions were fed into a Logistic Regression meta-classifier. Similarly, Horizon (185) conducted a broad comparison of different ensemble techniques, including AdaBoost, Bagging, and CatBoost. These methods operate on the principle that a committee of diverse models can often produce more stable and accurate predictions than any single model alone.

4 Results

The BeyondSmile Challenge yielded a rich set of results that reveal the strengths, weaknesses, and critical trade-offs of different computational strategies for depression detection. The performance of all eight participating teams under both the Universal (LOPO) and Hybrid (LOPDO) evaluation protocols provides a clear picture of the current state-of-the-art. A comprehensive summary of team methodologies and their corresponding performance is presented in Table 1.

To visualize the core findings, Figure 2 plots the performance of each team across both evaluation tracks. The figure highlights the trade-offs observed in the challenge: models that excelled in the generalized, universal setting did not necessarily perform well in the personalized, hybrid setting, and vice-versa. The following sections will analyze the performance within each track in detail.

4.1 Universal Model (LOPO) Performance

In the universal setting, which tests for pure generalization to unseen individuals, traditional machine learning pipelines built on extensive feature engineering demonstrated clear superiority. AI-MED Squad (142) achieved a perfect AUROC of 1.00 using a Random Forest classifier. Their strategy involved generating a vast array of statistical features from large, 6-hour time windows using libraries like tsfel and pycatch22. This result strongly suggests that by aggregating behavioral cues over long periods, it is possible to extract highly discriminative, "trait-like" features that generalize well across individuals. Other top performers in this track, such as Robot (118) achieved high AUROC scores of 0.94 by leveraging well-tuned classifiers on carefully selected or generated features.

Conversely, deep learning models that focused on learning from raw temporal sequences found the strict LOPO evaluation more challeng-

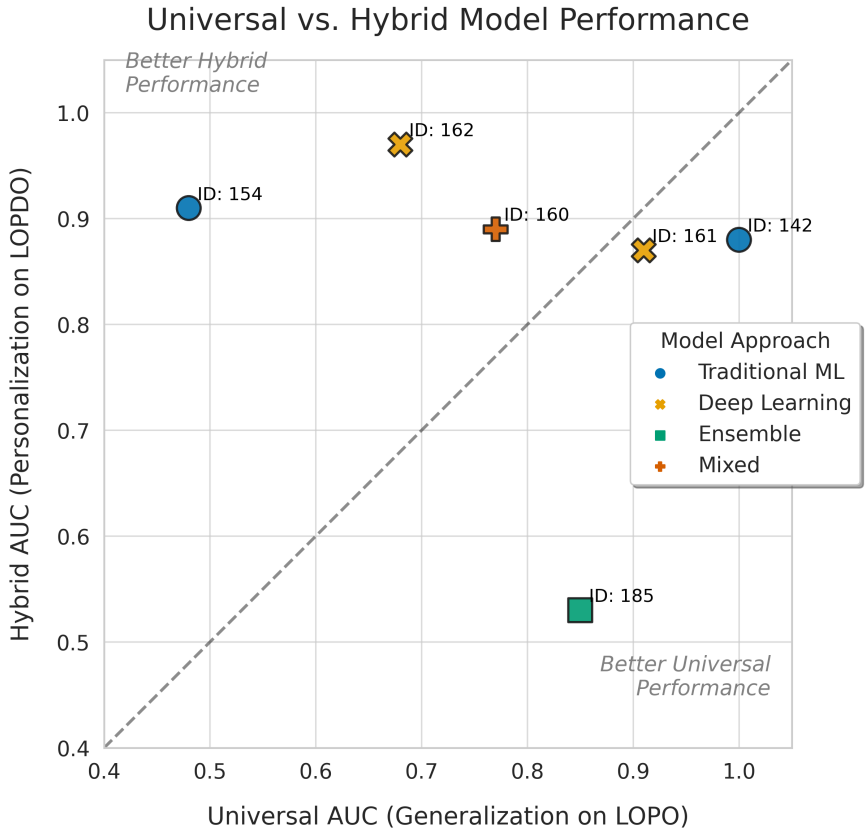

Figure 2: Universal vs. Hybrid Model Performance (AUROC). Each point represents a team's submission, colored by its primary methodological approach. The plot includes the 6 teams that submitted valid results for both the Universal and Hybrid evaluation tracks, allowing for a direct comparison (Team 118 reported only Universal scores, Team 119 reported neither). The dashed line (y=x) indicates equal performance. Points above the line excelled in the personalized (Hybrid/LOPDO) setting, while points below excelled in the generalized (Universal/LOPO) setting. The plot visually demonstrates the performance trade-off, with Deep Learning models clustering in the top-left and Traditional ML models in the bottom-right.

Table 1: Comprehensive Summary of Team Methodologies and Performance in the BeyondSmile Challenge

ID	Team Name	Approach	Key Feature Engineering Strategy	Win. Size	Univ. AUC	Hybrid AUC
142 [26]	AI-MED Squad	Trad. ML	Extensive statistical features (tsfel, pycatch22)	`6 hours`	**1.00**[b]	0.94
162 [33]	Lord of Depression	Trad. ML	Emotion/AU from contours of *first frame only*	`1 frame`	0.68	**0.97**
118 [11]	Robot	Trad. ML	Statistical features with explicit feature selection	`4 timesteps`	0.94	*N/A*
160 [13]	Team Persistence	Deep Learning	Sequential processing of comprehensive feature set	`1 timestep`	0.77	**0.88**
154 [21]	DP2K	Deep Learning	Geometric, statistical, and deep graphical features	`One video`	0.48	0.91
119[c] [25]	Team Access	Deep Learning	Raw AUs and classification probabilities in sequence	`3 timesteps`	N/A	N/A
161 [16]	Semicolon	Ensemble	Averaging of core AUs, head movement, smile/eye	`Full dataset`	0.91	0.87
185 [19]	Horizon	Mixed	Full features with and without imputation	`50 frames`	0.78	0.53

[a] It should be noted that teams employed different strategies for metric aggregation. Most results reflect a global (micro-averaged) metric calculated from all predictions, as per the challenge's guidelines. However, some submissions reported metrics as the average of per-participant scores, which can lead to variations in final values.
[b] This perfect score, while impressive, raises questions about potential overfitting due to the high-dimensional feature space relative to the number of samples, as discussed in Section V-D.
[c] Team 119 did not report any valid AUROC scores in their paper but was still accepted due to their efforts.

ing. DP2K (154), despite its sophisticated graphical Transformer model, scored a 0.48 AUROC. The eventual hybrid track winner, Team Persistence (160), scored a respectable 0.77 with their Bi-LSTM. This performance gap indicates that while sequence models are powerful, they can be more sensitive to the high inter-personal variance in facial behavior and may struggle to generalize across individuals without any user-specific context.

4.2 Hybrid Model (LOPDO) Performance

The performances changed dramatically in the hybrid evaluation setting, which allowed models to personalize predictions using a participant's past data. In this track, models capable of capturing temporal dynamics and adapting to user-specific patterns excelled.

Lord of Depression (162) achieved a near-perfect AUROC of 0.97. Their highly effective and counter-intuitive approach used an XGBoost model trained on AU and emotion features extracted from only the first frame of each recording. This suggests that in a personalized context, the model learns a user's baseline "resting" or "initial reaction" face, making deviations from this personal norm a powerful predictive signal. Deep learning models also showcased their strength in this personalized setting. The DP2K Transformer model's performance soared from 0.48 (LOPO) to 0.91 (LOPDO), validating its complex architecture's ability to learn intra-user patterns. The challenge winner, Team Persistence (160), achieved an outstanding 0.88 AUROC with their Bi-LSTM model, which effectively learned the sequential dependencies in each user's facial expressions and head movements over time. These results confirm that for personalized depression detection, modeling temporal dynamics is a more effective strategy than relying solely on aggregated statistical features.

4.3 Methodological Insights

Across both evaluation tracks, several key methodological lessons emerged:

- The "Window Size" Dilemma: The challenge revealed a significant lack of consensus on the optimal temporal window for feature extraction. Teams employed vastly different granularities, ranging from a single instantaneous frame (Lord of Depression), to short sequences of a few timesteps (Team Access, Robot), to large 6-hour aggregate blocks (AI-MED Squad). This variance underscores a critical methodological gap in the field and suggests that model performance is highly dependent on this choice of temporal aggregation.
- The Feature Engineering Trade-off: The results highlight a clear trade-off. For building universal, "one-size-fits-all" models, aggressive statistical feature engineering proved to be the winning strategy. For building personalized, adaptive models, focusing on learning from temporal sequences (even very short ones) was crucial.
- The Necessity of Rigorous Validation: The performance. As noted in the original summary draft, this was due to a fix for a data leakage issue. It powerfully illustrates that personalized models can easily

produce overly optimistic results if not evaluated with rigorous, time-series-aware nested cross-validation protocols like LOPDO.

5 Discussion

The BeyondSmile Challenge benchmarked a diverse set of computational strategies, and the results provide a nuanced understanding of the problem of depression detection in naturalistic settings. The starkly different outcomes between the Universal and Hybrid evaluation tracks are not just a matter of performance metrics; they reveal a fundamental dichotomy in what is being modeled: generalized behavioral traits versus personalized emotional states. This section discusses the key implications of our findings.

5.1 The Trait vs State Modeling Dichotomy

The most significant finding is the clear split in optimal strategies for the LOPO and LOPDO evaluations. This suggests that the two protocols are effectively measuring different aspects of depressive behavior.

- Universal Models Capture General "Traits": The success of AI-MED Squad (142) in the LOPO setting demonstrates that aggregating behavioral data over very long windows (6 hours) can produce powerful, user-agnostic biomarkers. By creating statistical features that summarize long periods, their model likely captured stable, "trait-like" characteristics of depression, such as a general reduction in facial expressivity or a persistent pattern of head movement, that are consistent across different individuals. This approach excels at creating a general-purpose screening tool but may miss subtle, day-to-day fluctuations.
- Hybrid Models Capture Personalized "States": In contrast, the LOPDO setting rewarded models that could capture transient, "state-like" changes in behavior. The success of the Bi-LSTM and Transformer models from Team Persistence (160) and DP2K (154) highlights the power of sequence modeling in a personalized context. These models are not just classifying a user; they are detecting deviations from that user's own established baseline. They excel at answering the question, "Is this person's behavior today indicative of a depressive state, given what I know about them from yesterday?"

The trend of hybrid models outperforming universal ones was not absolute, however. Some teams, like Horizon (185), observed a notable performance decrease in the hybrid setting (0.78 universal vs. 0.53 hybrid). This suggests that their chosen models, such as AdaBoost trained on aggregated features, were highly effective at capturing generalized, trait-like patterns but may have struggled to adapt to the more variable, high-frequency "state" data inherent to the day-level LOPDO protocol. This finding reinforces that the choice of model architecture must be closely aligned with the nature of the evaluation task, whether it prioritizes generalization or personalization.

5.2 Deep Learning Architectures

The challenge provided insights on the role of deep learning models like LSTMs and Transformers in this domain. While they struggled to generalize in the strict, user-agnostic LOPO setting (as evidenced by the low scores of DP2K (154)) their performance in the LOPDO setting.

This highlights a key trade-off. Deep sequence models possess the high capacity needed to learn the complex, non-linear temporal dependencies of an individual's facial dynamics. However, this same capacity can make them sensitive to the high inter-personal variance present in the dataset, leading to difficulties in generalizing to a completely new person without a "foothold" of prior data.

5.3 Methodological Gaps and the Path Forward

Finally, the challenge exposed several methodological gaps in the field. The "window size dilemma", where teams used temporal windows ranging from a single frame to six hours, indicates a lack of community consensus on how to process "in-the-wild" time-series data. This makes direct comparison of feature-based methods difficult and calls for future work on standardizing temporal processing.

5.4 The Challenge of Overfitting in High-Dimensional Feature Spaces

A crucial lesson from the challenge relates to the risk of model overfitting. The top-performing universal model from AI-MED Squad (142) achieved a perfect AUROC of 1.00. While an impressive result, it also raises important questions about model generalization, particularly when a "maximalist" feature engineering strategy is employed. Their approach generated thousands of features from a dataset with only a few dozen participant intervals, creating a high-dimensional space where finding a separating hyperplane might be possible even if the underlying patterns are not perfectly generalizable. This result highlights a critical consideration for future research: while aggressive feature engineering can unlock powerful predictive signals, it also carries a significant risk of overfitting to the specific characteristics of a small dataset. The true test of such a model would be its performance on a completely independent, held-out validation set, which was beyond the scope of this challenge. This underscores the need for the community to develop larger benchmark datasets to more robustly validate these powerful techniques.

6 Conclusion and Future Directions

The BeyondSmile Challenge has provided a valuable benchmark for the automated detection of depression from facial and head dynamics in naturalistic settings. While our analysis confirms the viability of automated facial behavior analysis for depression screening, it also highlights the critical importance of aligning model architecture with evaluation context

and the inherent risks of overfitting on small, high-dimensional datasets.

Our primary conclusion is the emergence of a clear "trait versus state" modeling paradigm. While complex deep learning models proved superior for personalized, state-based tracking, simpler ensemble models with extensive feature engineering paradoxically achieved near-perfect scores in the generalized, trait-based setting. This tension, along with the surprising efficacy of minimalist approaches in some contexts, provides a rich foundation for future work.

Based on the diverse submissions and our comprehensive analysis, we propose the following key directions for the research community to pursue:

- Standardizing Methodologies: The challenge exposed significant methodological variance, particularly in temporal windowing and metric calculation. Future work must aim to establish more standardized protocols to ensure that results across studies are robust and directly comparable.
- Validating on Larger, More Diverse Datasets: The perfect scores achieved by some models highlight an urgent need. To robustly validate high-capacity models and mitigate the risk of overfitting, future benchmarks must be conducted on larger, more diverse, multi-ethnic cohorts. This is essential for ensuring that models are learning true clinical patterns, not dataset-specific artifacts.
- Multimodal and Context-Aware Fusion: This challenge focused on visual cues, but a more holistic diagnostic picture requires more data. Future research should prioritize the fusion of facial data with vocal prosody, linguistic content, and other passively sensed behavioral data to build more accurate and contextually aware models.
- Prioritizing Explainable and Trustworthy AI (XAI): For any automated diagnostic tool to gain clinical acceptance, its decisions must be transparent. As demonstrated by several teams, incorporating interpretability methods like SHAP [20] is a crucial step. Future systems must be designed not just for accuracy, but also to provide clinicians and users with clear rationales for their predictions.

By addressing these future directions, the field can build upon the valuable insights gathered from this challenge and move closer to transitioning these promising research prototypes into trusted, accessible, and impactful clinical tools for mental health monitoring.

References

[1] Saeed Abdullah and Tanzeem Choudhury. Sensing technologies for monitoring serious mental illnesses. *IEEE MultiMedia*, 25(1):61–75, 2018.

[2] Constantino Álvarez Casado, Manuel Lage Cañellas, and Miguel Bordallo López. Depression recognition using remote photoplethysmography from facial videos. *IEEE Transactions on Affective Computing*, 2023.

[3] Prerna Chikersal, Afsaneh Doryab, Michael Tumminia, Daniella K Villalba, Janine M Dutcher, Xinwen Liu, Sheldon Cohen, Kasey G Creswell, Jennifer Mankoff, J David Creswell, et al. Detecting depres-

sion and predicting its onset using longitudinal symptoms captured by passive sensing: a machine learning approach with robust feature selection. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 28(1):1–41, 2021.

[4] M. Chowdhury, R. Sarmun, D. Bushnaq, M. Chabbouh, R. Aljindi, S. Pedersen, M. A. A. Faisal, N. Nahid, I. Hassan, R. Munemoto, C. Garcia, and S. Inoue. Summary of the silent speech decoding challenge using eeg data for arabic inner speech recognition. *International Journal of Activity and Behavior Computing*, 2025(2):1–21, 2025.

[5] Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre. Detecting depression from facial actions and vocal prosody. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7. IEEE, 2009.

[6] Sathyanarayanan Doraiswamy, Amit Abraham, Ravinder Mamtani, and Sohaila Cheema. Use of telehealth during the covid-19 pandemic: Scoping review. *J Med Internet Res*, 22(12):e24087, Dec 2020.

[7] Heiner Ellgring. *Non-verbal communication in depression*. Cambridge University Press, 2007.

[8] Denzil Ferreira, Vassilis Kostakos, and Anind K Dey. Aware: mobile context instrumentation framework. *Frontiers in ICT*, 2:6, 2015.

[9] Jeffrey M Girard, Jeffrey F Cohn, Mohammad H Mahoor, S Mohammad Mavadati, Zakia Hammal, and Dean P Rosenwald. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing*, 32(10):641–647, 2014.

[10] Jeffrey M Girard, Jeffrey F Cohn, Mohammad H Mahoor, Seyedmohammad Mavadati, and Dean P Rosenwald. Social risk and depression: Evidence from manual and automatic facial expression analysis. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2013.

[11] Bagus Hardiansyah, Fajar Astuti Hermawati, Dymas Adi Saputra, and Danara Dhasa Caesa. Depression detection via facial expressions and movement analysis using machine learning with optimized feature selection. *International Journal of Activity and Behavior Computing*, 2025(2):1–25, 2025.

[12] S. Hossain, T. Hossain, C. Garcia, T. Tazin, M. I. Mamun, T. Hujioka, S. Inoue, and M. A. R. Ahad. Understanding normal and unusual activities related to parkinson's disease from sensor data: Summary of activity and behavior computing tremor challenge. *International Journal of Activity and Behavior Computing*, 2025(2):1–18, 2025.

[13] Md. Rakibul Islam, Dip Sarker, Md. Tafhimul Haque Sadi, Labib Hasan Bayzid, Md. Kabiruzzaman, and Shahera Hossain. Facial behavior-based depression detection with bi-lstm and evaluation via universal and hybrid methods. In *2025 International Conference on Activity and Behavior Computing (ABC)*, pages 1–8, 2025.

[14] Rahul Islam and Sang Won Bae. Facepsy: An open-source affective mobile sensing system-analyzing facial behavior and head gesture for

depression detection in naturalistic settings. *Proceedings of the ACM on Human-Computer Interaction*, 8(MHCI):1–32, 2024.

[15] Rahul Islam and Sang Won Bae. Pupilsense: Detection of depressive episodes through pupillary response in the wild. *International Conference on Activity and Behavior Computing*, 2024.

[16] Aldo Januansyah, Rifa Amril Sahputra, Muhammad Adhiguna Hasibuan, Muhammad Yudya Ananda Hasibuan, Faiz Syukri Arta, and Muhammad Fikry. Optimizing facial expression and head dynamics data processing to enhance depression detection with cutting-edge ai models. In *2025 International Conference on Activity and Behavior Computing (ABC)*, pages 1–11, 2025.

[17] S Jennifer, Benjamin R Brady, Mohab M Ibrahim, Katherine E Herder, Jessica S Wallace, Alyssa R Padilla, and Todd W Vanderah. Co-occurrence of chronic pain and anxiety/depression symptoms in us adults: prevalence, functional impacts, and opportunities. *Pain*, 165(3):666–673, 2024.

[18] Xinru Kong, Yan Yao, Cuiying Wang, Yuangeng Wang, Jing Teng, and Xianghua Qi. Automatic identification of depression using facial images with deep convolutional neural network. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, 28:e936409–1, 2022.

[19] Yunus A. P. Chandraningtyas R. G. Pangestu H. G. Al Hauna A. D. Latief, M.A. and Y. Hong. Choo. Temporal-aware ensemble learning facial behavior analysis for accurate depression assessment. *International Journal of Activity and Behavior Computing*, 2025(2):1–25, 2025.

[20] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[21] Tu Nhat Khang Nguyen, Nguyen Khang Le, Mai Phuong Tran, Anh Duc Tran, and Nhat Tan Le. Geometrical, graphical, and transformer-based approaches for recognizing naturalistic depression in facial behavior. In *2025 International Conference on Activity and Behavior Computing (ABC)*, pages 1–9, 2025.

[22] Kennedy Opoku Asare, Isaac Moshe, Yannik Terhorst, Julio Vega, Simo Hosio, Harald Baumeister, Laura Pulkki-Råback, and Denzil Ferreira. Mood ratings and digital biomarkers from smartphone and wearable data differentiates and predicts depression status: A longitudinal data analysis. *Pervasive and Mobile Computing*, 83:101621, 2022.

[23] M. O'Brien and F. McNicholas. The use of telepsychiatry during covid-19 and beyond. *Irish Journal of Psychological Medicine*, 37(4):250–255, 2020.

[24] Paola Pedrelli, Szymon Fedor, Asma Ghandeharioun, Esther Howe, Dawn F Ionescu, Darian Bhathena, Lauren B Fisher, Cristina Cusin, Maren Nyer, Albert Yeung, et al. Monitoring changes in depression severity using wearable and mobile sensors. *Frontiers in psychiatry*, 11:584711, 2020.

[25] Sujay Saha, Maharshi Niloy, and Brototy Saha. Transformer-based depression detection model using facepsy data using facial action

units. *International Journal of Activity and Behavior Computing*, 2025(2):1–12, 2025.

[26] Justyna Skibińska, Abbas Shah Syed, Asma Channa, Zafi Sherhan Shah, Shehram Shah Syed, and Junaid Baloch. Recognising depression using facial and cephalic feature extractors with machine learning - answer to beyondsmile challenge. In *2025 International Conference on Activity and Behavior Computing (ABC)*, pages 1–11, 2025.

[27] Siyang Song, Shashank Jaiswal, Linlin Shen, and Michel Valstar. Spectral representation of behaviour primitives for depression analysis. *IEEE Transactions on Affective Computing*, 2020.

[28] Siyang Song, Shashank Jaiswal, Linlin Shen, and Michel Valstar. Spectral representation of behaviour primitives for depression analysis. *IEEE Transactions on Affective Computing*, 13(2):829–844, 2022.

[29] Anja Stuhrmann, Thomas Suslow, and Udo Dannlowski. Facial emotion processing in major depression: a systematic review of neuroimaging findings. *Biology of mood & anxiety disorders*, 1(1):1–17, 2011.

[30] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*, pages 3–10, 2014.

[31] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10, 2013.

[32] Sarah Collier Villaume, Shanting Chen, and Emma K Adam. Age disparities in prevalence of anxiety and depression among us adults during the covid-19 pandemic. *JAMA network open*, 6(11):e2345073–e2345073, 2023.

[33] Nguyen Phat Vo, Hoang Khang Phan, Anh Tuan Ha, Quang Le, and Nhat Tan Le. Eax: An emotion and action unit extraction framework for real time depression recognition. In *2025 International Conference on Activity and Behavior Computing (ABC)*, pages 1–9, 2025.

[34] Q. Xia, C. Garcia, U. Dobhal, X. Min, K. Tanigaki, N. Yoshimura, S. Inoue, T. Maekawa, and K. Wu. Summary of the virtual data generation for complex industrial activity recognition. *International Journal of Activity and Behavior Computing*, 2025(2):1–13, 2025.

Appendix

Participant Team Details

Table 2: Detailed Methodological and Computational Information for Participating Teams

ID	Team Name	Country	Machine Specification
142	AI-MED Squad	Poland, Pakistan, UAE	Intel CORE i7 - 1165G7 @ 2.80 GHz
162	LordofDepression	Vietnam	CPU: Apple M1 Pro, RAM: 32GB, GPU: Apple M1 Pro
118	Robot	Indonesia	Google Colab
160	Team Persistence	Bangladesh	Google Colab
154	DP2K	Vietnam	RAM: 128GB, CPU: Intel Xeon Silver, GPU: A100
119	Team Access	Bangladesh	Google Colab
161	Semicolon	Indonesia	Google Colabs v2-8 TPU
185	Horizon	Indonesia	RAM: 32GB, CPU: AMD Ryzen 7, GPU: NVidia GeForce RTX 4090