Don't Predict If You Cannot Interpret: Investigating the Clinical Viability of Facial Movements for machine learning Assisted Diagnostics of Bipolar Disorder

Martin Lund Trinhammer[1,5], Stella Graßhof*[1,5], Lars Vedel Kessing[2,4], Hanne Lie Kjærstad[4], Kamilla Woznica Miskowiak[3,4], and Sami S. Brandt[1,5]

[1]Data Science Section, IT University of Copenhagen, Copenhagen, Denmark

[2]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen

[3]Neurocognition and Emotion Across Disorders of the Brain (NEAD), Centre, Psychiatric Centre Copenhagen, Frederiksberg Hospital, Mental Health Services, Capital Region of Denmark

[4]Copenhagen Affective Disorder Research Centre (CADIC), Psychiatric Centre Copenhagen, Frederiksberg Hospital, Mental Health Services, Capital Region of Denmark

[5]Pioneer Centre for Artificial Intelligence, Copenhagen K, Denmark

June 9, 2026

Abstract

Background: Numerous studies have explored the possibility of developing automatic detection pipelines that can seamlessly diagnose patients with bipolar

*Corresponding author, stgr@itu.dk, Department of Computer Science, IT University of Copenhagen, Rued Langgards Vej 7, 2300 Copenhagen S, Denmark

disorder (BD) and other mental illnesses. Such novel diagnostic tools increasingly rely on data sources, such as facial movements, whose relationships to BD have yet to be fully elucidated. As such, these detection pipelines offer limited clinical value, despite promising performance estimates. A vital next step toward achieving clinically reliable models is to conduct granular interpretability analyses to determine which subsets of facial movements are responsible for determining patient or control class membership.

Materials and Methods: In this work, we rely on facial movements encoded as Action Units (AUs) of 32 participants recorded while watching emotional film clips. Our objective is to delineate the specific facial micro-movements responsible for the differences between patients with BD and controls by applying the interpretable Fisher's Linear Discriminant Analysis (LDA) in a binary, supervised classification design.

Results: We report how the movement of brow lowering (AU4) differentiates patients from controls with AUROC scores up to 69%.

Conclusions: Our exploratory study argues for the necessity of devising inherently interpretable machine learning models for the clinical domain. Furthermore, we critically discuss the implications of identifying AU4 as a key discriminative feature and assess the clinical value of specific facial movements for the diagnostic process.

Keywords— machine learning, bipolar disorder, interpretability, mental health

1 Introduction

Bipolar disorder (BD) is a serious mental illness affecting more than 1% of the global population [1]. In the International Classification of Diseases, 11th revision [2], BD is categorized as a mood disorder and characterized by alternating states of mood and energy, including episodes of mania or hypomania and depression, as well as mixed episodes. It is distinguished from unipolar depression by its biphasic nature. Many patients with BD also exhibit cognitive deficits and impairment, which are known to persist during episodes of euthymia, i.e., the absence of a mood episode [3].

Given its significant symptom overlap with unipolar depression, BD is commonly misdiagnosed and often first diagnosed several years after disease onset. This has severe consequences for the associated patient population, as the pharmacological and psychiatric treatment of BD differs markedly from that of unipolar depression. Early identification remains a vital ingredient in ensuring optimal treatment trajectories for the involved patients [4].

The current diagnostic process relies on a combination of psychometric inventories, such as the Hamilton Depression Scale (HRDS) and Young Mania Rating Scales (YMRS), and psychiatric assessment, an approach that is time-consuming to administer and lacks objectivity. To this end, substantial efforts have been undertaken to identify reliable diagnostic biomarkers

that could function as objective markers of the underlying mental state or disorder [5, 4].
One area associated with biomarker identification, which has seen a considerable surge in interest in the past years, is related to developing automatic detection systems, purportedly enabling the swift and early identification of BD [6, 7, 8]. Driven by developments in multimodal machine learning, a branch of Artificial Intelligence (AI) research, authors are increasingly exploring other data sources, such as facial movements, to aid in differential diagnosis. (For a review of machine learning assisted diagnosis using other modalities, we refer to [9, 10, 11]). Preliminary research using facial movement features suggests relatively high accuracy in identifying patient groups compared to control groups [12, 13, 9]. While these efforts remain interesting from a technical perspective, they currently lack the interpretative transparency to be considered viable for implementation. The main issue underlying the field of automatic detection of BD using facial movement features is that the differential patterns of facial movement between the patient and control groups are not fully accounted for by current approaches, which predominantly rely on "black-box" architectures [13, 12]. It is thus critical to analyze the precise micro-movements responsible for differences in facial patterns between patients with BD and controls. Predictive screening tools, which are based on a poor or nonexistent understanding of how features relate to a disease, are essentially unreliable for implementation in clinical practice because the decision logic remains opaque. The model may be using a shortcut or inferring bias from its training data, which is very difficult to determine as long as the relevance of its features is unclear.
Our objective in the current volume is to address these shortcomings in research by offering a thorough investigation into facial movement patterns of patients with BD using an intrinsically interpretable framework. To this end, we utilize a subset of the data reported in [14], which employs an emotional film clip viewing paradigm. The goal is to investigate if it is possible to identify specific and reliable facial markers that enable the demarcation of the BD patient group. Efforts like these are vital to establishing the clinical foundation underpinning the future of automatic detection systems, which have the potential to strengthen the current diagnostic process.

1.1 Interpretability of computational methods for diagnostics using facial movements

While multiple studies have explored the detection of BD from facial movements, our central claim is that such methods must be statistically interpretable to gain relevance in a clinical context. In this section, we further develop this argument and define interpretability within the scope of the present study.
From an interpretability perspective, prior work on differentiating patients from controls using facial movements falls into three distinct categories. First, the most significant volume of studies identified rely on neural networks for classification and predominantly use raw pixel values from images and videos as input [12, 15, 9, 16]. These models are notoriously difficult

to interpret [17] given that their complex architectures function as "black boxes", yielding high classification accuracy but offering little intrinsic insight into the underlying features. A second cluster of contributions applies various post-hoc techniques to gauge individual-pixel importance, such as Gradient-weighted Class Activation Mapping (Grad-CAM) and saliency maps [18, 19]. Such approaches are generally helpful in ensuring that the model has not learned a spurious relationship in the input data; however, the results obtained across multiple images or videos are difficult to aggregate and disseminate to the broader research community. A third set of studies uses feature-extraction tools, such as OpenFace [20], combined with machine learning methods. Interpretability is predominantly achieved here via Shapley Additive Explanations (SHAP scores) [21], a powerful tool for gauging individual feature importance. While SHAP is a practical, model-agnostic tool, we argue that its necessity often stems from the choice of an initially opaque model architecture. Consistent with recent arguments in explainable AI research, we believe it is more appropriate to devise a model architecture that is intrinsically interpretable, rendering the need for post-hoc explainability techniques unnecessary [22, 23, 24, 25, 26]. In the following, we will show how such an inherently interpretable architecture can be devised using a simple linear model and a feature-penalization framework. Crucially, this will allow us to examine the performance of the final model and to ascertain whether it has learned a clinically valid relationship.

2 Material and Methods

2.1 Participants

Participants for the current study were recruited as part of a larger Danish cohort study (Bipolar Illness Onset) that was explicitly designed to investigate biomarker identification for BD. The Bipolar Illness Onset study has been approved by the Local Ethical Committee (H-7-2014-007) and the data agency, Capital Region of Copenhagen (RHP-2015–023), and adheres to the Declaration of Helsinki principles [5]. The data used in the present study were initially described in [14], which is the only other study in which it appears. The present research uses a subset of the original dataset, comprising 16 patients with BD and 16 healthy controls (HC). This subset was choosen due to the availability of an audio track only for the selected participants, which we require for temporal alignment of the recordings (further details in 2.2.1). All participants gave informed consent before participation. Patients were recruited if they were in complete or partial remission, indicating no current manic or depressive episode. This was psychometrically identified using the HDRS and YMRS, with scores of 14 or lower included for each inventory. As is observed in Table 3, the groups were well-matched for age ($p = 0.85$), gender ($p = 0.48$), and years of education ($p = 0.30$), indicating no significant demographic differences between the cohorts. Clinically, the BD group exhibited a mean HDRS-17 score of 6.56 ($SD = 4.35$) and a mean YMRS of 2.81 ($SD = 2.90$). While significantly higher than controls, these scores remain below the standard

Table 1: Description of film clips adopted from [14].

Movie	Description
Neutral	A man explains how to make a cup of tea while making one himself.
Happy	A mother sits with her four laughing babies. The laughter continues for more than a minute.
Sad	Footage of the Holocaust.
Winning	The Danish swimmer Pernille Blume wins a gold medal at the Olympic Games 2016, while the sports commentators cheer enthusiastically.
Racing	Two cars are racing; at the very last minute, they both avoid crashing into a train.
Risk-taking	A person without safety equipment walks around and jumps onto the roof of a skyscraper.
Socially anxious	A socially awkward situation where a blindfolded woman farts in front of her friends at her own surprise party, thinking she is alone.

clinical cutoffs for active episodes, confirming that the BD participants were in a state of euthymia or remission during the study. Patients also underwent psychiatric assessment to confirm diagnosis and absence of a current acute mood episode, which was evaluated using the Schedule for Clinical Assessment in Neuropsychiatry (SCAN) interview undertaken by a medical or psychological PhD student [27]. Matching HCs were included if they could report no psychiatric illness either personally or within first-degree family. The experimental setup invited participants to watch seven film clips, while their spontaneous facial movements were recorded. The seven movie clips were initially chosen to elicit specific emotional responses to BD, allowing the identification of distinct expression patterns among the two groups of participants, and are described in Table 1. The videos were presented to the participants in random order and lasted for a total of 13 minutes.

2.2 Preprocessing

Our complete processing pipeline is visualized in Figure 1. We now turn to a thorough description of each step performed.

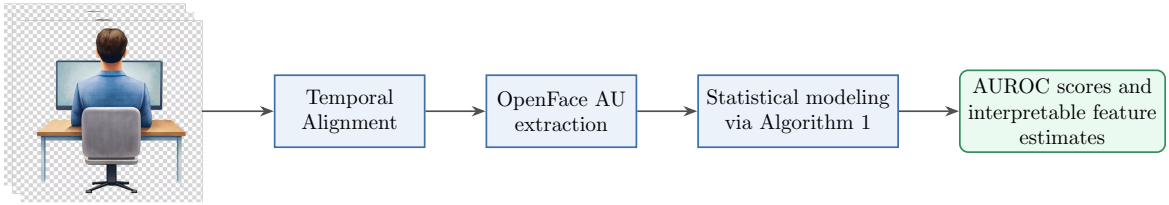

Figure 1: Data preprocessing pipeline. Please refer to Algorithm 1 for details of the modeling.

2.2.1 Temporal alignment

As mentioned, the current study works with a subset of the data reported in [14]. This subset was selected because an audio track was available for specific participants, enabling high-

quality temporal video alignment. Alignment is a critical step in our analysis, as we observed diverging starting times between participant recordings and the stimulus video's activation. We would thus not know when the stimulus video began for each recording, rendering analysis of the participants' facial movements and comparison to the group unfeasible. Aligning the recordings by audio track alleviates this issue, which is why audio-track availability was an inclusion criterion for our study. In total, 224 recordings are identified for analysis, with 32 participants each viewing seven different movie clips. For each of these movie clips, we aligned the tapes using the audio track as follows. First, we observe a slight linear shift in the recording start times. That is expected, as the recording might have started slightly before the stimulus video began. We computed the cross-correlation between the stimulus videos and the participants' corresponding recordings. The highest value is reached where the subsequences are most similar, i.e., have the most significant overlap, and hence provides a plausible estimate of the time delay between the recordings. We then removed a few excess frames at the beginning and cropped the recordings to the same length, ensuring temporal consistency across recordings for the same stimulus clip. We show the number of frames available per recording in Table 4.

2.2.2 Action unit extraction

We utilize the well-established OpenFace toolkit [20] for estimation of the intensity of 16 facial Action Units (AUs) per frame [28]. Each AU encodes a specific muscle movement of the face, as outlined in Table 2. The operationalization of facial movements according to AUs is the most commonly used method in the fields of computer vision and affective computing, and is inspired by Carl-Herman Hjortsjö [29], with subsequent development and formulation by Paul Ekman and Wallace Friesen [28]. We selected the intensity outputs for the AUs in question (range 0.00-5.00) and mean-corrected these on a per-movie basis. This last normalization step ensures the data conveys genuine facial movements rather than participants' constant facial structure or other spurious features. We intentionally focus exclusively on facial movements —considering only AUs —and thus exclude estimates for gaze and head pose estimation.

2.3 Statistical analyses

The analyses of our study are divided into three subsections: 1) an explorative data analysis, 2) predictive modeling, and 3) participant subgrouping.

2.3.1 Exploratory data analysis

To build initial intuition about group differences, we performed an exploratory analysis of the raw (non-mean-corrected) data across all feature combinations. Here, we visualize the mean signal response for the BD and control groups, averaged across all 7 movie conditions. We include the figure on the computation for AU4 in our results section.

Table 2: Facial Action Unit (AU) codes and their associated facial motion from [20].

AU	Facial movement
AU1	Inner brow raiser
AU2	Outer brow raiser
AU4	Brow lowerer
AU5	Upper lid raiser
AU6	Cheek raiser
AU7	Lid tightener
AU9	Nose wrinkler
AU10	Upper lip raiser
AU12	Lip corner puller
AU14	Dimpler
AU15	Lip corner depressor
AU17	Chin raiser
AU20	Lip stretched
AU23	Lip tightener
AU25	Lips part
AU26	Jaw drop

2.3.2 Predictive modeling with LDA

To address the predictive performance and generalizability of the AUs in discriminating between patients and controls, we implemented a supervised binary classification experiment using Linear Discriminant Analysis (LDA) [30] (See section 8 for details). LDA was chosen for its strong performance and interpretability [31], as it classifies samples by projecting the data onto a space that maximizes class separation. Recognizing that the inherent multicollinearity among AUs can obscure true feature importance in standard linear models, we employed a robust, nested feature selection strategy utilizing the Least Angle Regression (LARS) regularization path [32]. Within each cross-validation fold, the LARS algorithm was first applied to the training data to generate a complete regularization path, representing a sequence of models of increasing complexity. At every step along this path, the resulting subset of AUs was used to train an LDA model, and its performance was evaluated on an independent validation set. The specific feature set that yielded the highest classification performance on the validation set was designated as the optimal set for that fold. Finally, an LDA model was retrained on the complete training set of the outer fold using this optimal feature set, and its predictive performance was evaluated on the independent test set, providing an unbiased measure of generalization and reliable insight into the predictive magnitude of the selected features.

Interpretability by sparsity and transparency Using the LARS algorithm for feature selection prior to an LDA model ensures interpretability by combining sparsity with the geometric transparency of LDA. LARS shrinks the coefficients of irrelevant or redundant

features, preventing multicollinearity dimensionality issues that can render an LDA model's coefficients unreliable. By reducing the input to a high-signal subset of variables, the LDA is able to fit a robust linear decision boundary where the remaining coefficients are straightforward to interpret. The magnitude of each coefficient provides a measure of relative importance, while the sign indicates the directional influence on class separation.

Algorithm The specific training procedure for the experiments described in 2.3.2 has been conducted using stratified, nested cross-validation and is formulated as a binary classification problem. The objective is supervised, binary classification of BD or control class membership using the AU intensity estimates as inputs. The pipeline is outlined in Algorithm 1. For the outer loop, we utilize 16 folds (F_{outer}). This maximizes the training data available from 32 participants, leaving 30 participants for training and validation, with one case and one control retained for each test evaluation. For the inner loop, we use a 15-fold split (F_{inner}). We observed how the order of the participants had a small effect on the results. To account for this, we run the pipeline described in Algorithm 1 N_{trials} times. By examining the standard deviation of the performance metrics across consecutive runs, we found that $N_{\text{trials}} = 12$ was the minimum required to ensure stable, consistent results. The experiments are thus evaluated across $N_{\text{trials}} = 12$ distinct executions of the pipeline, with the participant vector order varying each time. The results reported in Table 5 are the mean over $Nfolds(16) * Ntrials(12) = 192$ test-folds. All experiments have been evaluated using the AUROC metric. Since our input data from the OpenFace preprocessing is structured by AU levels *per frame*, we must aggregate this information to obtain a single score per person, since the labels are provided at the per-person level. This is achieved through our frame aggregation method described in 9. The core benefit of this approach, compared to a simple mean-aggregation, is how it accounts for the inherent correlation between frames. This results in one probability estimate per person, which we use as input to the AUROC score calculation, together with the actual label. All experiments have been conducted in Python 3, using sklearn for model implementations [33]. Claude Opus 4.5 has been used for occasional code review.

2.3.3 Participant sub grouping

In the third and final part of our experiments, we investigate differences in how participants are classified for each specific film clip. This is included to elucidate opportunities for subgrouping patients and is shown in Figure 3.

3 Results

3.1 Sample characteristics

Our sample comprises a fully balanced dataset with 16 patients and 16 HCs; see Table 3. As described in 2.2.2, for each participant, we extract AUs per frame corresponding to the

Algorithm 1 Complete pipeline for LDA and LARS + LDA

Input: N_{AU} AUs per frame for all participants
Output: Mean AUROC over trials on the independent test set

Perform mean correction on the data
for trial $t = 1$ to N_{trials} **do**
  Shuffle order of participants
  Split participants into F_{outer} stratified folds, keeping one case and one control in each fold.
  for each movie m **do**
    Test Set: Hold out the participants in the test fold
    Split the remaining participants into F_{inner} stratified folds for training and validation
    LDA
    Train LDA on the full training and validation set
    Evaluate AUROC on the held-out test set
    LARS + LDA
    for each inner split s (training and validation) **do**
      Feature selection path: Apply LARS to the inner training set to generate the regularization path (sequence of feature sets) [32]
      for each feature set k along the LARS path **do**
        Train LDA using feature set k on the inner training set
        Compute AUROC on the inner validation set (AUROC_k)
      end for
      Define $\text{Features}_{\text{Optimal}}$ as the set k yielding $\max(\text{AUROC}_k)$
    end for
    Final Model: Train LDA using $\text{Features}_{\text{Optimal}}$ on the full training/validation set
    Evaluate AUROC on the held-out Test Set
  end for
end for
Report the mean over all trials and folds per movie for both the LDA and LARS + LDA conditions.

aligned recordings. Table 4 summarizes the number of frames available both per recording and in total, resulting in 16 AU estimates for 641.952 frames.

3.2 Exploratory data analysis: AUs

The first point of our investigation concerns the primary data analysis of the AUs. First, we report that the BD group shows a higher AU4 intensity than control participants, as evidenced by the markedly higher means in Figure 2. This indicates that the movement of brow lowering underpinning AU4 is more pronounced for the patient group. We further report Mann-Whitney U tests for significance for the top four AUs across our 7 distinct movie conditions (see supplementary material 10 and Table 6). Despite the limited statistical power in individual movie conditions ($N = 32$), AU4 demonstrated remarkable consistency, ranking

Table 3: Summary of participant information for healthy controls and patients.

	Bipolar disorder Mean (SD)	**Healthy controls** Mean (SD)	**Statistics** (U/χ^2)	p **(2-tailed)**
N	16	16		
Age	32.69 (11.06)	33.81 (11.71)	123.00	0.85
Gender (% female)	10 (63%)	8 (50%)	0.51	0.48
Education, years	14.63 (2.78)	15.50 (2.92)	100.50	0.30
HDRS-17	6.56 (4.35)	1.00 (1.21)	20.50	$< .001$
YMRS	2.81 (2.90)	.94 (1.18)	75.00	0.047
BD type (% BD-II)	11 (69%)			
Illness duration, years[a]	7.69 (8.59)			
Untreated BD, years[b]	6.75 (8.58)			
Antidepressants, no (%)	1 (6%)			
Antipsychotic, no (%)	5 (31%)			
Anticonvulsants, no (%)	6 (38%)			
Lithium, no (%)	2 (13%)			

Table 4: Number of frames available per movie condition.

Movie	**Number of Frames**	**x**	**Participants**	**=**	**Total number of frames**
Neutral	3013	x	32	=	96.412
Happy	2620	x	32	=	83.840
Sad	2822	x	32	=	90.304
Winning	2750	x	32	=	88.000
Racing	3170	x	32	=	101.440
Risk taking	2868	x	32	=	91.776
Socially anxious	2818	x	32	=	90.176
Total				=	**641.952**

as the top discriminative feature in 5 out of 7 conditions and second in the remaining two. In all cases, AU4 achieved near-significance levels ($p \leq .10$), a stability unmatched by any other AU.

3.3 Predictive modeling: Classifier performance with LDA

Following the exploratory analysis, we investigated the predictive performance of the feature set. We have evaluated our experiments over three different conditions: first, using all 16 AUs as inputs to the LDA model (All Features); second, only using AU4 (AU4); and third, using LARS for feature selection across all 16 AUs (LARS + LDA) [32] with an LDA model performing the classification. We report how relying on AU4 yields the highest performance with an AUROC score of 69% for the movie conditions "Sad" and "Socially anxious". While the LARS feature selection should, in theory, yield at least as good performance as the AU4

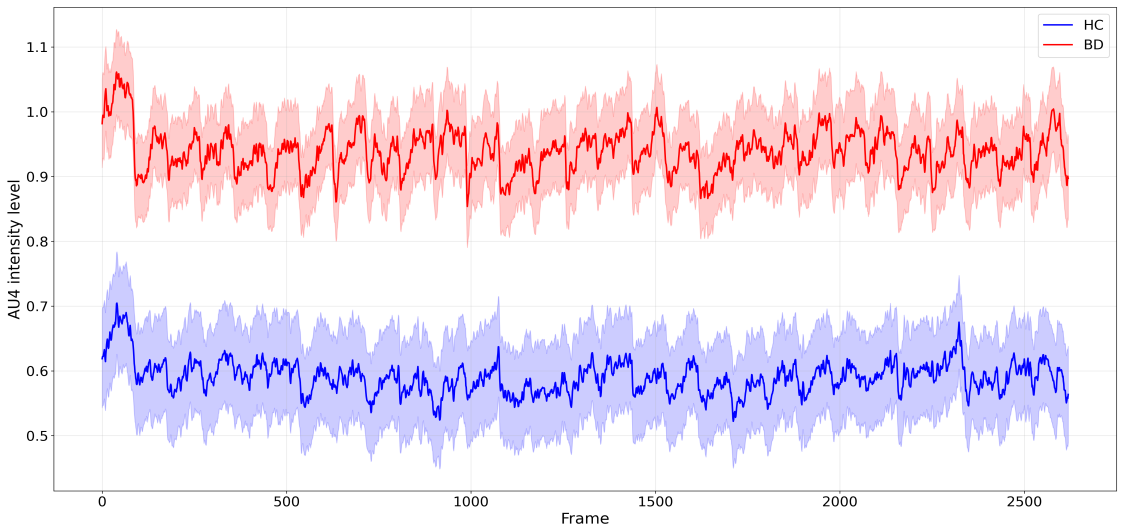

Figure 2: Mean differences between the BD patients and controls for AU4 across the seven movie clips, non-mean corrected data. Truncated to the shortest movie sequence.

Table 5: Comparison of AUROC scores using LDA with different feature sets (mean over 192 folds). The parentheses next to the LARS + LDA condition designate the share of folds in which AU4 was selected as the first feature, regardless of subsequent features.

Movie	**All Features**	**Only AU4**	**LARS + LDA**
Happy	0.59	0.66	0.56 (82%)
Sad	0.46	**0.69**	0.47 (85%)
Neutral	0.49	0.61	0.49 (90%)
Winning	0.54	0.67	0.55 (95%)
Risk taking	0.54	0.68	**0.63** (98%)
Socially anxious	**0.60**	**0.69**	0.61 (100%)
Racing	0.45	0.60	0.36 (77%)
Mean	**0.52**	**0.66**	**0.53**

condition, this is not the case in our dataset, as the regularization overfits to certain noisy features. This happens even though LARS selects AU4 as the first, and thus most prominent, feature in the vast majority of folds. In the "Socially anxious" movie condition, AU4 is selected as the first feature in all folds; however, this does not yield the same performance as the sole AU4 condition, as the LARS algorithm also selects other features to combine with AU4, which then fail to generalize as well to the test samples. To confirm the robustness of our chosen approach relying on LARS for feature selection, we executed the same pipeline using Recursive Feature Elimination (RFE) for feature selection [34], instead of LARS. These results are shown in 11 and are moderate compared to those obtained with LARS + LDA. This is likely due to the RFE search pattern relying on the LDA model's feature ranking; as we argued previously, one cannot naively use LDA feature weights to gauge importance because

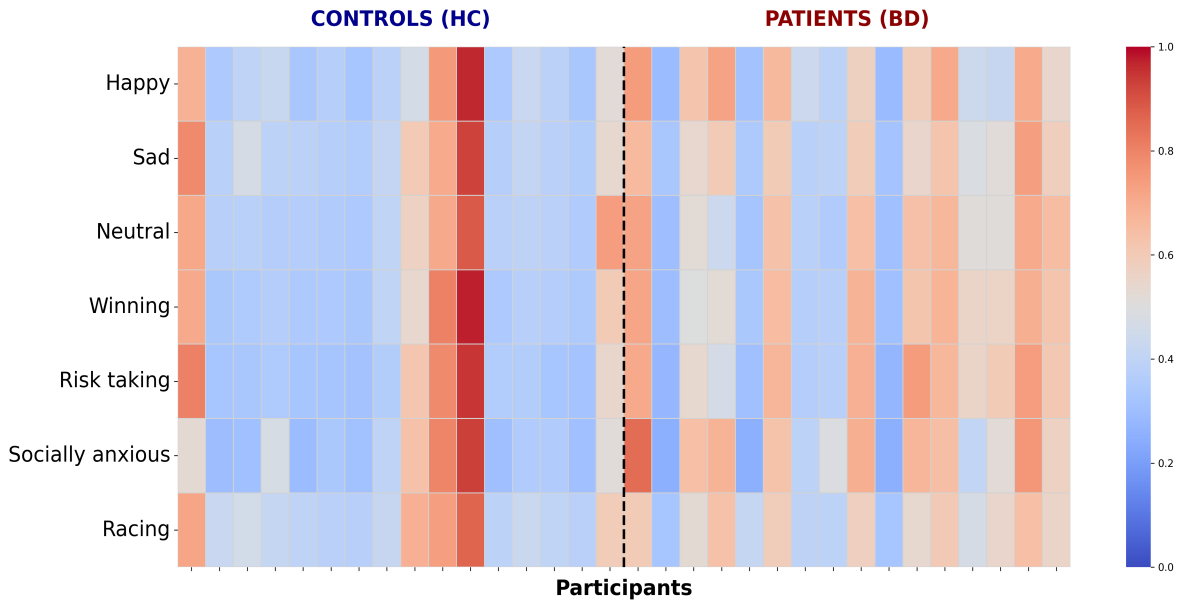

Figure 3: Predicted AUROC probability per participant per movie achieved by executing Algorithm 1 only using AU4 for the LDA condition. Red values indicate probabilities closer to 1 (i.e., the patient group). Each tick represents a participant, with controls on the left-hand side and BD on the right. The Y axis denotes each movie condition. Mean probability per person across 12 random seeds.

of issues with multicollinearity. We refer to our supplementary material 12 for an exhaustive investigation of the potential confounding effect of gender. This experiment confirms the absence of a consistent gender signal in AU4, indicating that gender is not a confounding variable in our study.

3.4 Participant subgrouping for AU4

Given the robust and consistent performance of the LDA model using only AU4 as the input feature, we select this condition for further analysis of how each participant is classified across the movie conditions. In sum, we execute our Algorithm 1 for the LDA-condition (no LARS) with only AU4 as input and report the AUROC scores across all participants. These findings are summarized in Figure 3, which displays the predicted probabilities for control and BD patients. We observe a trend towards alignment in the vertical plane; i.e., the predicted probabilities are consistent within participants but vary more across participants. The findings suggest that the specific stimulus in the movie condition has a lesser impact on participants' facial movements.

4 Discussion

4.1 Summary of findings

In this study, we investigated differences in facial movement patterns between patients with BD and control participants during emotional film-clip viewing. A core objective of ours is to present a modeling framework that enables robust feature interpretability, as this is critical for achieving clinical transparency in machine learning assisted diagnostics. Using an LDA model together with the LARS algorithm, we have found that brow lowering (AU4) is markedly more pronounced in the patient group than in the HCs. When using AU4 as the sole feature input to a predictive LDA classifier, the BD patient group can be distinguished from HCs with an AUROC of up to 69%. By coupling the sparse feature selection capabilities of the LARS algorithm with the geometric interpretability of LDA, we establish a transparent modeling pipeline that allows us to move beyond the opacity often associated with high-dimensional modeling. The results of this investigation unequivocally identify AU4 as the dominant feature driving the group differences. Upon deeper examination of specific group and within-group characteristics, we observe how the specific film clip viewed by the participants appears less salient in driving group differences. With a maximum score of 69%, our results are aligned with current benchmarks for facial micro movements for BD [13], and for the depressive spectrum in general [9, 12]. These results are achieved through the application of a straightforward-to-interpret model, a quality vital for domains such as mental health. This further increases the clinical usefulness of our results and also underscores the value and performance of explainable architectures.

4.2 Brow lowering (AU4)

As shown in Table 5, our best overall performance is achieved solely by relying on AU4. While an overall negative bias in facial affect, also known as blunted affect, is an established finding among patients with serious mental illness [35], our findings suggest the possibility that AU4 may drive this negativity bias. This trend likely stems from the fact that most studies investigating this negativity bias focus on anger, sadness, and fear. According to the Facial Action Coding Scheme (FACS), the operationalization of all three emotions requires AU4 [28]. Our findings thus critically underscore the need for future studies to analyze which AUs drive the computation of prototypical emotion categories, as a single feature may skew the measures. Suppose AU4 skews the estimates of sadness, anger, and fear. In that case, this may not be visible to many researchers applying proprietary software to extract facial information, as the setup of, for instance, the Noldus FaceReader [36] or iMotions [37] easily enables the user to see the prototypical emotion categories, rather than which features are driving the differences.

The possibility of increased brow lowering being a transdiagnostic marker of mood disor-

ders should be considered. This hypothesis is closely related to how the common treatment of glabellar botulinum toxin injections is proposed to function [38]. Both a common aesthetic treatment and one increasingly associated with positive outcomes for mood disorders [39, 40], the treatment involves injecting botulinum toxin into the frown lines of the face. The positive effect observed in patients with mood disorders is assumed to function by altering the proprioceptive feedback mechanism [41, 42, 43], which posits a relationship between the body's physical appearance and the experience of emotions. Botulinum toxin helps relax facial muscles and produces a more positive facial expression, which is then associated with more positive emotions. These considerations help us place our findings in a relevant context, thereby suggesting an appropriate interpretation of why AU4 is salient among BD patients. Further studies would need to confirm the robustness of this finding before the interpretation can be confirmed.

Our investigation into the per-participant scores on the AU4 dimensions further reveals the potential of this measure for subgrouping patients. The predicted probabilities per person and stimulus video are illustrated in Figure 3. Efforts to identify subgroups within the current clusters of mental illnesses have been suggested by multiple authors, due to the high symptom heterogeneity within each illness category [44, 45]. The opportunity to identify such patient subgroups is critical for developing personalized treatment, which holds the potential to improve patient outcomes. As the results from Figure 3 suggest, the degree of brow lowering is consistent for each participant across the movie conditions; it appears that the specific emotional stimulus from the movie clip holds less value in distinguishing the groups. While the results from Table 5 suggest more pronounced differences according to the movie condition, these differences are slightly inflated due to the use of the AUROC metric with only two participants in each test-evaluation. Additionally, in Figure 3, we observe that our model confidently classifies a control participant as belonging to the patient group, as shown in the eleventh column, where the probability estimates exceed 90%. A deeper investigation of this outlier revealed that the participant's facial morphology naturally exhibits very pronounced glabellar lines, which likely led OpenFace to critically overestimate AU4 activation. If we consider related studies from other disciplines that have shown prominence of AU4, findings from computerized education settings stand out. With efforts to predict engagement and learning outcomes among university students, [46] reports how high levels of AU4 functioned as a negative predictor of endurability and a positive predictor of frustration. As shown in Figure 2, the BD patient group showed markedly higher AU4 levels than control participants. By leaning on findings from computerized education settings, it cannot be ruled out that the high AU4 levels among BD patients may be an artifact of the computerized study design, in which the patient group is more likely to experience frustration while watching the movie clips. An explanation for this may lie in the well-established findings of cognitive impairment and emotion dysregulation among patients with mood disorders, including BD [47, 48, 49]. Following this logic, deficits in executive functioning and/or emotion dysregulation among this study's sample of BD patients may make these participants more likely to be confused

or emotionally agitated than control participants, which, in turn, may drive the heightened AU4 levels.

4.3 Clinical implications

As our results demonstrate, group differences can be identified through facial movement analyses, underscoring the technical feasibility of diagnostic screening and detection tools built on facial features. Its value from a clinical perspective, however, hinges on the ability to know for sure that the relationship identified by the model is appropriate to the domain. While we argue that the finding of heightened levels of AU4 is in line with other findings on facial movements from individuals with BD [14], the current literature does not render it possible to explain with absolute certainty why this relationship occurs. Such inference is needed before deployment is clinically appropriate, and likely depends on a combination of both theoretical developments from the clinical domain and new model architectures tailored accordingly [50]. It should further be acknowledged that the predictive performance of many suggested detection pipelines is generally higher when a multi-modal setup is used, i.e., when also text-based or vocal features are included [13, 12]. These studies, however, often fail to analyze the specific cross-modal feature combinations that drive predictive performance, thereby severely limiting clinical transparency and reliability. The use of deep learning models usually makes it impractical to extract feature importance estimates. The present findings suggest how an interpretable model architecture can perform on par with deep learning models. Future data fusion techniques, which are necessary to merge data from different modalities, should be developed with explainability as a core objective, thereby enabling the identification of perhaps unknown patterns in patients' behavior. Additionally, it is essential for future studies to also account for differential diagnostics, taking into account the possibility of patients having multiple disorders and also comparing different clinical groups with one another. Such endeavors are often stymied by data availability, though efforts are mounting to address these issues [51].

4.4 Strengths and limitations

A notable strength of our research design is its reliance on a robust temporal alignment framework, which enables high-quality frame alignment via cross-correlation. This ensures we can confidently compare participants on a per-frame basis. This comes with the limitation of a smaller sample size, as only a subset of participants had an audio track required for alignment. However, our sample of 32 participants is equally divided between the patient and control groups ($n = 16$ each). While we acknowledge a minor difference in gender distribution, the patient group is clinically coherent, as all BD patients are in either full or partial remission. Second, as previously indicated, the finding of marked increases in brow lowering has a strong theoretical link to the proposed mechanism of action of botulinum toxin treatment. As this treatment is commonly applied to patients of unipolar depression (UD), it can be speculated

that the finding of increased brow lowering is more a feature characterizing UD. It is thus a limitation of our research design that we lack a UD comparison group. Third, while the experimental paradigm of emotional film-clip viewing has the advantage of ensuring the same stimuli for each participant, the associated facial movements may not generalize outside the computerized setting [52]. Fourth, because we rely on data extraction from OpenFace [20], we are limited to considering facial movement patterns categorized as AUs. If participants, for instance, touch their face in various ways during the experimental setting, this is not reflected in our features, even though such gestures may hold clinical value. Finally, we must address the ethical implications of automated diagnostics based on facial movement features. A key factor to consider is the involuntary nature of facial movements. In contrast to data originating from questionnaires, facial micro-expressions are conveyed without conscious awareness or control, rendering them particularly sensitive. There is a risk that such technologies could be applied to non-clinical population surveillance, potentially to restrict access to societal resources (for instance, with regard to insurance or employment). Further, facial movement data represents highly sensitive biometric data that cannot be effectively anonymized. To mitigate these risks, such systems should function solely as 'decision-support' tools—never replacing human judgment—and must be deployed only with the patient's explicit, informed consent, with full transparency regarding the model's interpretability.

5 Conclusion

In this study, we have demonstrated that the facial micro-movement associated with brow lowering (AU4) can distinguish BD patients from controls, with an AUROC of up to 69%, which is in line with related benchmarks, but with a vastly more interpretable modeling architecture. Our results further suggest that future studies involving facial affective display in computerized settings should pay close attention to the intensity of individual AUs, rather than relying on the prototypical emotion categories, as these categories are easily skewed by a subset of or a single feature. We strongly encourage future studies developing automatic detection pipelines and digital phenotyping for mental illnesses to report the features that drive the model's performance, as the research community needs to evaluate and develop the theoretical underpinnings of the relationship identified by the model.

6 Acknowledgment

This project was partially funded by the Pioneer Centre for AI, DNRF grant number P1

7 Declaration of Interest Statement

HLK reports a relationship with Lundbeck LLC that includes: consulting or advisory. LVK reports a relationship with Lundbeck LLC that includes: consulting or advisory. LVK reports

a relationship with Teva Pharmaceutical Industries Ltd that includes: consulting or advisory. KWM reports a relationship with Lundbeck LLC that includes: consulting or advisory. KWM reports a relationship with Gideon Richter that includes: consulting or advisory. KWM reports a relationship with Angelini Pharma Inc that includes: consulting or advisory. KWM reports a relationship with Janssen-Cilag Ltd that includes: consulting or advisory. Other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Eduard Vieta, Michael Berk, Thomas G. Schulze, André F. Carvalho, Trisha Suppes, Joseph R. Calabrese, Keming Gao, Kamilla W. Miskowiak, and Iria Grande. Bipolar disorders. *Nature Reviews Disease Primers*, 4(11):1–16, March 2018.

[2] World Health Organization. International classification of diseases, 11th revision (icd-11). https://icd.who.int, 2019.

[3] André F Carvalho, Beatrice Bortolato, Kamilla Miskowiak, Eduard Vieta, and Cristiano Köhler. Cognitive dysfunction in bipolar disorder and schizophrenia: a systematic review of meta-analyses. *Neuropsychiatric Disease and Treatment*, page 3111, December 2015.

[4] Eduard Vieta, Estela Salagre, Iria Grande, André F. Carvalho, Brisa S. Fernandes, Michael Berk, Boris Birmaher, Mauricio Tohen, and Trisha Suppes. Early intervention in bipolar disorder. *American Journal of Psychiatry*, 175(5):411–426, May 2018.

[5] Lars Vedel Kessing, Klaus Munkholm, Maria Faurholt-Jepsen, Kamilla Woznica Miskowiak, Lars Bo Nielsen, Ruth Frikke-Schmidt, Claus Ekstrøm, Ole Winther, Bente Klarlund Pedersen, Henrik Enghusen Poulsen, Roger S. McIntyre, Flavio Kapczinski, Wagner F. Gattaz, Jakob Bardram, Mads Frost, Oscar Mayora, Gitte Moos Knudsen, Mary Phillips, and Maj Vinberg. The bipolar illness onset study: research protocol for the bio cohort study. *BMJ Open*, 7(6):e015462, June 2017.

[6] Shuang Gao, Vince D. Calhoun, and Jing Sui. Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neuroscience Therapeutics*, 24(11):1037–1052, 2018.

[7] Ma Ruihua, Zhao Meng, Chen Nan, Liu Panqi, Liu Sijia, Zhao Ke, Tan Shuping, Tian Li, and Wang Zhiren. Differences in facial expression recognition between unipolar and bipolar depression. *Frontiers in Psychology*, 12:619368, 2021.

[8] Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(11):1–14, November 2022.

[9] Zifan Jiang, Salman Seyedi, Emily Griner, Ahmed Abbasi, Ali Bahrami Rad, Hyeokhyen Kwon, Robert O Cotes, and Gari D Clifford. Multimodal mental health digital biomarker analysis from remote interviews using facial, vocal, linguistic, and cardiovascular patterns. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[10] Giovanni Briganti and Jérôme R. Lechien. Voice quality as digital biomarker in bipolar disorder: A systematic review. *Journal of Voice*, January 2025.

[11] Rongrong Zhong, XiaoHui Wu, Jun Chen, and Yiru Fang. Using digital phenotyping to discriminate unipolar depression and bipolar disorder: Systematic review. *Journal*

of Medical Internet Research, 27(1):e72229, May 2025. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research publisher: JMIR Publications Inc., Toronto, Canada.

[12] Lang He, Mingyue Niu, Prayag Tiwari, Pekka Marttinen, Rui Su, Jiewei Jiang, Chenguang Guo, Hongyu Wang, Songtao Ding, Zhongmin Wang, et al. Deep learning for depression recognition with audiovisual cues: A review. *Information Fusion*, 80:56–86, 2022.

[13] Francesco Ceccarelli and Marwa Mahmoud. Multimodal temporal machine learning for bipolar disorder and depression recognition. *Pattern Analysis and Applications*, 25(3):493–504, August 2022.

[14] Hanne Lie Kjærstad, Caroline Kamp Jørgensen, Ingrid Broch-Due, Lars Vedel Kessing, and Kamilla Miskowiak. Eye gaze and facial displays of emotion during emotional film clips in remitted patients with bipolar disorder. *European Psychiatry: the journal of the Association of European Psychiatrists*, 63(1), 2019.

[15] 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–e936409–12, July 2022.

[16] Ghulam Gilanie, Sana Cheema, Akkasha Latif, Anum Saher, Muhammad Ahsan, Hafeez Ullah, and Diya Oommen. A robust method of bipolar mental illness detection from facial micro expressions using machine learning methods. *Intelligent Automation & Soft Computing*, 39(1), 2024.

[17] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85, 2022.

[18] Jonathan Chung, Moshe Eizenman, Uros Rakita, Roger McIntyre, and Peter Giacobbe. Learning differences between visual scanning patterns can disambiguate bipolar and unipolar patients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[19] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[20] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2018.

[21] Pınar Baki, Heysem Kaya, Elvan Çiftçi, Hüseyin Güleç, and Albert Ali Salah. A multimodal approach for mania level prediction in bipolar disorder. *IEEE Transactions on Affective Computing*, 13(4):2119–2131, 2022.

[22] C Rudin. Stop explaining black box models for high stakes decisions and use interpretable models instead. nature machine intelligence, 1 (5), 206–215, 2019.

[23] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[24] José Pereira Amorim, Pedro Henriques Abreu, João Santos, and Henning Müller. Evaluating post-hoc interpretability with intrinsic interpretability. *arXiv preprint arXiv:2305.03002*, 2023.

[25] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The lancet digital health*, 3(11):e745–e750, 2021.

[26] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

[27] John Kenneth Wing, Thomas Babor, TS Brugha, J Burke, John E Cooper, Rob Giel, Assen Jablenski, Darrel Regier, and Norman Sartorius. Scan: schedules fonr clinical assessment in neuropsychiatry. *Archives of general psychiatry*, 47(6):589–593, 1990.

[28] Paul Ekman, Wallace V Friesen, and JC Hager. Facial action coding system: manual. palo alto. *CA: Consulting Psychologists*, 1978.

[29] Carl-Herman Hjortsjo. Man's face and mimic language. *Studen litteratur.*, 1969.

[30] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006.

[31] Qing Zhao, Hong-Zhen Fan, Yan-Li Li, Lei Liu, Ya-Xue Wu, Yan-Li Zhao, Zhan-Xiao Tian, Zhi-Ren Wang, Yun-Long Tan, and Shu-Ping Tan. Vocal acoustic features as potential biomarkers for identifying/diagnosing depression: a cross-sectional study. *Frontiers in Psychiatry*, 13:815678, 2022.

[32] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. 2004.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[34] Xue-wen Chen and Jong Cheol Jeong. Enhanced recursive feature elimination. In *Sixth international conference on machine learning and applications (ICMLA 2007)*, pages 429–435. IEEE, 2007.

[35] Tovah Cowan, Michael D Masucci, Tina Gupta, Claudia M Haase, Gregory P Strauss, and Alex S Cohen. Computerized analysis of facial expressions in serious mental illness. *Schizophrenia research*, 241:44–51, 2022.

[36] Elisa Landmann. I can see how you feel—methodological considerations and handling of noldus's facereader software for emotion measurement. *Technological Forecasting and Social Change*, 197:122889, 2023.

[37] iMotions A/S. *iMotions (9.3)*. Copenhagen, Denmark, 2022.

[38] Tillmann HC Kruger and M Axel Wollmer. Depression–an emerging indication for botulinum toxin treatment. *Toxicon*, 107:154–157, 2015.

[39] Marc Axel Wollmer, Michelle Magid, Tillmann HC Kruger, and Eric Finzi. Treatment of depression with botulinum toxin. *Toxins*, 14(6):383, 2022.

[40] M Magid, E Finzi, THC Kruger, HT Robertson, BH Keeling, S Jung, JS Reichenberg, NE Rosenthal, and MA Wollmer. Treating depression with botulinum toxin: a pooled analysis of randomized controlled trials. *Pharmacopsychiatry*, 25(06):205–210, 2015.

[41] Randy J Larsen, Margaret Kasimatis, and Kurt Frey. Facilitating the furrowed brow: An unobtrusive test of the facial feedback hypothesis applied to unpleasant affect. *Cognition and emotion*, 6(5):321–338, 1992.

[42] Pamela K Adelmann and Robert B Zajonc. Facial efference and the experience of emotion. *Annual review of psychology*, 40(1):249–280, 1989.

[43] Marc Heckmann, Bianca Teichmann, Ulrike Schröder, Reiner Sprengelmeyer, and Andrés O Ceballos-Baumann. Pharmacologic denervation of frown muscles enhances baseline expression of happiness and decreases baseline expression of anger, sadness, and fear. *Journal of the American Academy of Dermatology*, 49(2):213–216, 2003.

[44] Eiko I Fried and Randolph M Nesse. Depression is not a consistent syndrome: An investigation of unique symptom patterns in the star* d study. *Journal of affective disorders*, 172:96–102, 2015.

[45] Yen-Ling Chen, Pei-Chi Tu, Tzu-Hsuan Huang, Ya-Mei Bai, Tung-Ping Su, Mu-Hong Chen, and Yu-Te Wu. Identifying subtypes of bipolar disorder based on clinical and neurobiological characteristics. *Scientific reports*, 11(1):17082, 2021.

[46] Joseph Grafsgaard, Joseph B Wiggins, Kristy Elizabeth Boyer, Eric N Wiebe, and James Lester. Automatically recognizing facial expression: Predicting engagement and frustration. In *Educational data mining 2013*, 2013.

[47] Kamilla W Miskowiak and Cecilia S Petersen. Neuronal underpinnings of cognitive impairment and-improvement in mood disorders. *CNS spectrums*, 24(1):30–53, 2019.

[48] Kamilla Woznica Miskowiak, KE Burdick, A Martinez-Aran, CM Bonnin, CR Bowie, AF Carvalho, P Gallagher, B Lafer, C López-Jaramillo, T Sumiyoshi, et al. Assessing and addressing cognitive impairment in bipolar disorder: the international society for bipolar disorders targeting cognition task force recommendations for clinicians. *Bipolar disorders*, 20(3):184–194, 2018.

[49] Hanne Lie Kjærstad, Julian Macoveanu, Gitte Moos Knudsen, Sophia Frangou, K Luan Phan, Maj Vinberg, Lars Vedel Kessing, and Kamilla Woznica Miskowiak. Neural responses during down-regulation of negative emotion in patients with recently diagnosed bipolar disorder and their unaffected relatives. *Psychological Medicine*, 53(4):1254–1265, 2023.

[50] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Manish Raghavan. The inversion problem: Why algorithms should infer mental state and not just predict behavior. *Perspectives on Psychological Science*, 19(5):827–838, 2024.

[51] Ziad Obermeyer and Sendhil Mullainathan. Health data platforms. Technical report, National Bureau of Economic Research, 2022.

[52] Lisa Feldman Barrett. Context reconsidered: Complex signal ensembles, relational meaning, and population thinking in psychological science. *American Psychologist*, 77(8):894, 2022.

8 Methodological Details

In this appendix, we will outline the mathematical and implementation details of our machine learning methods.

Let $\mathbf{X}$ be the $N \times M$ data matrix, where each of the M columns represents one AU, and each row captures the varying AU intensities over all frames concatenated over all participants. The data matrix $\mathbf{X}$ is accompanied by a binary label vector $\mathbf{y}$ indicating if the corresponding frame belongs to the patient or control.

8.1 Two-stage approach

For the results reported in Table 5 (column 3: LARS + LDA), we employ LARS as a feature selection mechanism to identify a sparse, discriminative subset of AUs, which is subsequently used to train an LDA classifier.

8.1.1 LARS

LARS provides an efficient mechanism to identify a sparse, discriminative subset of AUs by computing the entire regularization path. Given the training data $(\mathbf{X}, \mathbf{y})$ with M features, LARS produces a sequence of features $\mathcal{F} = (f_1, f_2, \ldots, f_M)$ which are ranked according to their significance. This ordering reflects the strength of the features' association to the labels, where f_1 is the most predictive feature. This defines a sequence of feature subsets:

$$\mathcal{S}_1 = \{f_1\} \subset \mathcal{S}_2 = \{f_1, f_2\} \subset \cdots \subset \mathcal{S}_M = \{f_1, \ldots, f_M\}. \tag{1}$$

For each candidate subset an LDA classifier is trained on the inner training fold and evaluated on the inner validation fold.

8.1.2 Linear Discriminant Analysis

To establish core notation: a sample $\mathbf{x}$ is assigned a label as follows

$$\text{if } y = \mathbf{w}^T\mathbf{x} \geq 0 \text{ control}, \tag{2}$$

$$\text{if } y = \mathbf{w}^T\mathbf{x} < 0 \text{ patient}. \tag{3}$$

To retrieve the parameter vector $\mathbf{w}$ in the LDA terminology, the objective is to maximize the distance between the means of the two classes, while minimizing the within-class scatter matrix. Hence, we aim to maximize the objective function

$$f(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}, \tag{4}$$

where $\mathbf{S}_B$ denotes the between-class scatter matrix, defined as:

$$\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T. \tag{5}$$

Here, $\mathbf{m}_1$ and $\mathbf{m}_2$ are the mean of the data points in classes C_1 and C_2. The within-class scatter matrix $\mathbf{S}_W$ is defined as:

$$\mathbf{S}_W = \sum_{n \in C_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^T + \sum_{n \in C_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^T, \tag{6}$$

where $\mathbf{x}_n$ represents the data point belonging to either class C_1 or C_2. The means $\mathbf{m}_1$ and $\mathbf{m}_2$ are calculated accordingly:

$$\mathbf{m}_1 = \frac{1}{N_1} \sum_{n \in C_1} \mathbf{x}_n, \quad \mathbf{m}_2 = \frac{1}{N_2} \sum_{n \in C_2} \mathbf{x}_n,$$

where N_1 and N_2 denote the number of frames in each class. Following this, the class probabilities are calculated through maximum likelihood estimation. We rely on default hyperparameters as defined in sklearn.

9 Aggregate frames to a per-person predicted probability

As mentioned in Algorithm 1, our method is designed to classify each frame. Thus, a solution is needed to aggregate the probabilities per frame to a unified posterior probability per person. This method needs to take into account the correlation between frames. To obtain one posterior probability per person, which can be fed to the AUROC metric for evaluation, we use the following frame aggregation method.

Let P_t be the estimated probability for frame t to belong to the BD patient class, computed according to LDA independently for each frame. We define the modified likelihood as

$$\tilde{L}(q, c) = \left(\prod P_t^q (1 - P_t)^{1-q} \right)^c = \; L(q)^c, \tag{7}$$

where the class label is $q \in \{0, 1\}$ and $c \in [0, 1]$ is a complexity parameter modeling the dependence between frames. The parameter posterior distribution is thus,

$$p(q, c \mid D) = \frac{\tilde{L}(q, c)\, P(q)}{\sum_{q=0}^{1} \tilde{L}(q, c)\, P(q)}, \tag{8}$$

where $P(q)$ is the prior probability of the class, assuming a uniform prior for c. Taking c as

a nuisance parameter, and D as the data, the marginal posterior probability is

$$P(q \mid D) = \int_0^1 p(q, c \mid D)\, dc \propto P(q) \int_0^1 L(q)^c\, dc = \left(\frac{L(q) - 1}{\log L(q)} \right) P(q). \tag{9}$$

Assuming number of frames is an order of magnitude larger than 1, we have

$$P(q \mid D) \approx \frac{(\log L(q))^{-1} P(q)}{\sum_{q=0}^{1} (\log L(q))^{-1} P(q)}, \tag{10}$$

where for the balanced case $P(0) = P(1) = 0.5$. Since $\log L(q)^{-1}$ is the code length of the original class probabilities, the posterior probability can be interpreted as the relative code length of the true BD and control data sequences.

10 Significance tests for all AUs

In Table 6, we present significance tests of the most discriminating AUs for each movie condition. We intentionally choose only to report the top four AUs, to avoid showing redundant information. While the limited sample size of individual film clips reduced the statistical power to detect effects at the $p < .05$ level, AU4 consistently emerged as the primary differentiator between the clinical and control groups. AU4 ranked as the most significant feature in 5 out of 7 conditions and second in the others. The fact that AU4 is consistently among the top-ranking features confirms our previous results and highlights it as a marker of BD, distinguishing it from other facial movements which appeared only sporadically.

Table 6: Top 4 AUs by statistical difference (Mann-Whitney U) between patients of BD and controls across movie conditions.

Happy

Action Unit	*U*	*p*
AU04	84.0	0.101
AU09	153.0	0.356
AU10	152.0	0.376
AU15	104.0	0.376

Sad

Action Unit	*U*	*p*
AU04	82.0	0.086
AU01	156.0	0.300
AU17	101.0	0.318
AU25	105.0	0.396

Neutral

Action Unit	*U*	*p*
AU23	93.0	0.194
AU04	94.0	0.207
AU12	95.5	0.221
AU17	100.0	0.300

Winning

Action Unit	*U*	*p*
AU04	78.0	0.062
AU14	101.0	0.318
AU10	153.0	0.356
AU05	149.0	0.440

Risk Taking

Action Unit	*U*	*p*
AU04	82.0	0.086
AU25	97.0	0.250
AU15	150.0	0.418
AU14	108.0	0.462

Socially Anxious

Action Unit	*U*	*p*
AU04	80.0	0.073
AU09	162.0	0.207
AU07	155.0	0.318
AU10	145.0	0.534

Racing

Action Unit	*U*	*p*
AU12	96.0	0.223
AU04	97.0	0.250
AU26	102.0	0.337
AU05	152.0	0.376

11 Robustness check: Recursive Feature Elimination

To bolster the robustness of our LARS feature selection method, with results reported in Table 5, we evaluated a secondary feature selection pipeline using RFE with an LDA classifier. The results shown in Table 7 were computed by executing our Algorithm 1 with the same specifications as reported in our LARS + LDA condition in Table 5, however, using RFE for the feature selection, instead of LARS. A key distinction between the two approaches is the feature selection pattern: LARS operates as a forward selection algorithm, computing the entire selection path in a single efficient computation. In contrast, RFE is a backward elimination algorithm that begins with the full feature set and iteratively removes the least informative features. Most importantly, the ranking in the RFE pipeline is determined by the LDA weights. The lower performance reported in Table 7 for our RFE experiments stems from the LDA model performing the feature ranking rather than the LARS coefficients. As

mentioned in Section 2.3.2, LDA suffers from multicollinearity, meaning the model cannot determine exact feature importance when features are correlated.

Table 7: Comparison of AUROC scores using LDA with RFE with different feature sets (mean over 192 folds). The parenthesis next to RFE + LDA designate the share of folds in which AU4 was selected as the first feature, regardless of subsequent features.

Movie	**RFE + LDA**
Happy	0.49 (32%)
Sad	0.40 (21%)
Neutral	0.40 (29%)
Winning	0.53 (52%)
Risk Taking	0.56 (92%)
Socially Anxious	0.59 (78%)
Racing	0.36 (7%)
Mean	**0.48**

12 Gender confounding check

To ensure that the clinical differences reported in Table 5 were not confounded by gender, we conducted a series of Mann-Whitney U tests comparing male and female participants across all AUs and movie conditions, reported in Table 8. Crucially, AU4 did not appear among the top four discriminating features for gender in any of the seven conditions. While AU4 was consistently shown to be a primary differentiator between clinical groups (as observed in Table 6), it shows no statistical signal for gender differentiation. This lack of association confirms that AU4 is specific to the clinical group in our sample and is not confounded by gender.

Table 8: Top 4 AUs by statistical difference (Mann-Whitney U) between male and female across movie conditions.

Happy			**Sad**		
Action Unit	*U*	*p*	*Action Unit*	*U*	*p*
AU10	86.0	0.133	AU02	184.0	0.029
AU12	88.0	0.154	AU07	85.0	0.124
AU06	88.0	0.154	AU06	104.0	0.410
AU09	164.0	0.154	AU17	104.0	0.414
Neutral			**Winning**		
Action Unit	*U*	*p*	*Action Unit*	*U*	*p*
AU12	53.5	0.005	AU12	78.0	0.070
AU06	64.5	0.020	AU06	81.5	0.094
AU10	66.5	0.024	AU07	84.0	0.115
AU07	79.0	0.077	AU26	92.0	0.203
Risk Taking			**Socially Anxious**		
Action Unit	*U*	*p*	*Action Unit*	*U*	*p*
AU06	77.0	0.063	AU15	196.0	0.008
AU12	79.5	0.078	AU17	193.0	0.012
AU07	82.0	0.098	AU23	181.0	0.038
AU10	85.0	0.124	AU25	177.0	0.055
Racing					
Action Unit	*U*	*p*			
AU07	86.0	0.133			
AU20	146.0	0.459			
AU06	111.5	0.592			
AU05	139.0	0.635			