# SHAP-based Prediction of Mother's History of Depression to Understand the Influence on Child Behavior

Maneesh Bilalpur
Saurabh Hinduja
mab@pitt.edu
University of Pittsburgh
Pittsburgh, Pennsylvania, USA

Laura Cariola
The University of Edinburgh
Edinburgh, United Kingdom

Lisa Sheeber
Oregon Research Institute
Eugene, Oregon, USA

Nicholas Allen
University of Oregon
Eugene, Oregon, USA

Louis-Philippe Morency
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Jeffrey F. Cohn
University of Pittsburgh
Pittsburgh, Pennsylvania, USA

## ABSTRACT

Depression strongly impacts parents' behavior. Does parents' depression strongly affect the behavior of their children as well? To investigate this question, we compared dyadic interactions between 73 depressed and 75 non-depressed mothers and their adolescent child. Families were of low income and 84% were white. Child behavior was measured from audio-video recordings using manual annotation of verbal and nonverbal behavior by expert coders and by multimodal computational measures of facial expression, face and head dynamics, prosody, speech behavior, and linguistics. For both sets of measures, we used Support Vector Machines. For computational measures, we investigated the relative contribution of single versus multiple modalities using a novel approach to SHapley Additive exPlanations (SHAP). Computational measures outperformed manual ratings by human experts. Among individual computational measures, prosody was the most informative. SHAP reduction resulted in a four-fold decrease in the number of features and highest performance (77% accuracy; positive and negative agreements at 75% and 76%, respectively). These findings suggest that maternal depression strongly impacts the behavior of adolescent children; differences are most revealed in prosody; multimodal features together with SHAP reduction are most powerful.

## CCS CONCEPTS

• **Applied computing** → **Psychology**; • **Human-centered computing**; • **Computing methodologies** → **Feature selection**;

## KEYWORDS

adolescent behavior, depression, feature selection, multimodal prediction

**ACM Reference Format:**
Maneesh Bilalpur, Saurabh Hinduja, Laura Cariola, Lisa Sheeber, Nicholas Allen, Louis-Philippe Morency, and Jeffrey F. Cohn. 2023. SHAP-based Prediction of Mother's History of Depression to Understand the Influence on Child Behavior. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 9–13, 2023, Paris, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3577190.3614136


*ICMI '23, October 9–13, 2023, Paris, France*

ACM ISBN 979-8-4007-0055-2/23/10…$15.00
https://doi.org/10.1145/3577190.3614136

## 1 INTRODUCTION

Depression strongly influences the interpersonal behavior of affected individuals. Some examples of altered behavior include attenuated head motion and facial movements [19]. Depressed individuals are perceived to be sadder, more negative, and more uncomfortable [7]. This altered interpersonal behavior may have significant consequences for families with children.

About 10-15% of mothers with minor children are affected by depression each year [10]. Developmental studies [24] find that mothers with depression express increased negative affect with their children. Such differences in parenting influence children's development [15]. Effects include impaired social outcomes and cognitive functioning and increased occurrence of internalizing and externalizing problems (such as depression and social anxiety). Little is known, however, about how depression in mothers influences the proximal behavior of their children. In an effort to reveal possible transmission of depression to offspring, we investigated the discernible characteristics of children of depressed mothers as compared to their counterparts.

Much of what we know from developmental studies of mother-child interactions is based on observer ratings of behavior. One example of an observational coding system to study behavior in family environments is the Living in Family Environments (LIFE) system [17]. LIFE codes distinguish between positive, aggressive, dysphoric and neutral behavior. Such coding approaches provide high-level, subjective descriptors but are unable to capture objective behavioral differences within and between behavioral modalities.

To capture low-level behavior, computational measures are needed. These measures been actively used for depression detection and severity prediction in clinical interviews and non-family contexts but not within families. We use computational measures of behavior to understand differences in child behavior related to maternal depression. We evaluate the hypothesis that children of depressed and non-depressed mothers differ strongly between and within modalities of behavior. We compare unimodal behavior to multimodal behavior in predicting maternal depression from child behavior,

early vs. late fusion strategies, and introduce a novel feature selection strategy based on the SHapley Additive exPlanation framework (SHAP)[20]. In addition, we evaluate the efficacy of computational measures by comparing them with manually annotated observer ratings of behavior.

SHapley Additive exPlanations (SHAP) [20] of features has been used to understand how specific features affect prediction models [3]. SHAP-based feature selection approaches typically entail inclusion of additional random or shadow features [21, 32]. We propose an approach to SHAP ranking that avoids the need to introduce shadow or random features and significantly reduces the number of features. These innovations help in both being conservative with the number of features as well as incorporating explainability into the solution. The novel contributions of this work include the following:

(1) Detecting children of mothers with a history of depression from multimodal computational measures of behavior.
(2) Comparing computational measures and manual annotation by expert human raters.
(3) Using SHapley Additive exPlanations (SHAP) for feature selection without introducing masked or random features.
(4) Identifying the relative contribution of modalities and features within modalities that vary between children of depressed and non-depressed mothers.

## 2 RELATED WORK

Previous work has focused on occurrence and to lesser extent history of depression in adults. To the best of our knowledge, our work is the first to address influence of depression using computational measures in children of parents with versus without depression. For that reason, related works necessarily are limited to those that predict current or history of depression in adults rather than in children of affected families.

Generalization of features across datasets for current depression prediction was recently explored by Alghowinem et al. [2]. They explored various feature selection methods using an SVM classifier to study how summary statistics of multimodal features generalize across datasets in three different countries. Using features capturing various non-verbal cues (head dynamics and gaze), speech behavior and prosody, this work highlighted the challenges in feature generalization due to differences in the nature of datasets. For example, features from BlackDog (patient-therapist interview) and AVEC (interaction with a computer) datasets found generalization on the Pitt dataset (patient-therapist interview) challenging. They found that a subset of features from gaze and prosody generalized across datasets.

Alternatives to patient-therapist interactions for depression prediction have also been explored. One such example is the audio-visual data collected through a virtual agent as interviewer in the DAIC-WoZ dataset [16]. More recently, Ray et al. [28] used audio, text and video features from the virtual agent driven conversation to extract both and deep learning based and interpretable features. These included sentence embeddings for text, eGeMAPs features along with embeddings of audio signal from the DeepSpectrum deep learning model, and pose, gaze and action unit features from videos. They proposed using the attention mechanism both within and between modalities and showed large improvement with respect to the baseline. Interpreting the predictions has been limited to modality-level understanding that text features were more important (higher attention weight) while both audio-visual features were given less but equal importance in the prediction.

Baki et al. [3] used features from prosody, verbal, and action units modalities for predicting levels of mania in persons with bipolar disorder. SHAP explanations for decision fusion among unimodal features indicate that formant frequencies, loudness, jitter and shimmer are some of the most contributing features in prosody. Lip corner puller (AU 12) and chin raiser (AU 17) from action units, and religious talk and linguistic modality are also some top predictors for mania level in bipolar disorder.

Following the preliminary study by Cariola et al. [6] on linguistic differences between mothers with and without history of depression, Bilalpur et al. [5] predicted history of depression among mothers. They used verbal and non-verbal (without prosody) behavioral features extracted from mothers for the prediction task. By eliminating collinear features and using 15 features from different modalities, a 78% separability was reported. Among many other correlations with depression, they found that depressed mothers tend to be less analytic, use more negations and comparisons, smile with less intensity and talk less.

Barring differences in the nature of the problem, most recent work fails to include one or more modalities that may contribute to depression detection. Bilalpur et al. [5] omitted prosody in studying the history of depression among mothers. Alghowinem et al. [2] limited features to prosody, turn dynamics (referred to as speech behavior) and dynamics of the face and head. Ray et al. [28] omitted effect turn dynamics. To understand multimodal behavior, we use a comprehensive set of verbal and non-verbal features that were not studied in conjunction before. Our feature set consists of facial action units, face and head dynamics, speech behavior, prosody, and linguistics.

To identify the relative contribution of features within and between modalities, we use a SHAP-based ranking approach. SHAP for feature selection was earlier explored by Marcilio et al. [21]. In the process of selection, they introduced masked features to replace the existing ones while respecting the rank order. Later PowerSHAP [32] was proposed that introduced one random feature into the existing set of features. The feature selection is performed based on the intuition that an informative feature on average has a higher mean absolute SHAP value over a random feature.

Existing SHAP-based feature selection methods either preserve the number of features through masking or increase them by introducing random features in the process of feature selection. Our proposed approach in subsection 3.5 focuses on both decreasing the number of features and increasing the performance.

## 3 DATASET AND METHODS

### 3.1 Dataset

The mother-child dyadic interaction dataset from Nelson et al. [23, 24] was used in this work. Mothers and their adolescent child were involved in a 15-minute problem-solving interaction. Depressed dyads were recruited based on the mother's history of

treatment for depression and elevated current depressive symptoms. A non-depressed comparison group consisted of dyads with mothers with no history of treatment for depression and no current symptomatology. 84% of participants were White; minorities were distributed across American Indian/Alaskan, Native Hawaiian/Pacific Islander, African American or multiple ethnicity. Age range in children was limited to early adolescence, and all families were low-income. Audio was captured at 16kHz and video at 30fps with dedicated hardware for mother and child. Though 152 interactions were transcribed, 4 of them were excluded due to the lack of manual observational codes (see subsection 3.3). This resulted in 148 dyads. 73 dyads had a mother with history of treatment for depression.

## 3.2 Multimodal Features

To derive multimodal features, we used various computational frameworks available for affective computing. The resultant features comprised Action Units (AU), Face & Head dynamics, Speech Behavior, Linguistic and Prosody modalities. Features were limited to those from the child. The constituents of these modalities along with how they were used in the classification framework are discussed below.

*Action Units.* The occurrence and intensity of facial muscle movements based on the anatomically-driven FACS (Facial Action Coding System) [9] were included in the analysis. Both intensity and occurrence dimensions predicted from the AFARtoolbox [11, 25] were used to operationalize duration. These were then summarized through mean, minimum, maximum, standard deviation, interquartile range and variance statistics. Both variance and standard deviation were included as features to account for their non-linear relationship, which would escape a linear classifier (in this work, SVM with a linear kernel). To quantify duration, occurrence likelihoods of AUs were thresholded at 0.5 to define the presence of an AU. Because duration of the dyadic interaction tasks could differ, duration of continuous AU occurrence was normalized by the duration of the dyadic interaction task. A similar approach was followed for AU intensity. This resulted in 156 features.

(1) Intensity for AUs 6, 10, 12, 14, 6+12 (positive affect)
(2) Likelihood of occurrence for AUs 1, 2, 4, 7, 15, 17, 23, 24
(3) Duration of occurrence for AUs 1, 2, 4, 6, 7, 10, 12, 6+12, 14, 15, 17, 23, 24

*Face & Head Dynamics.* Face and head dynamics primarily consisted of facial landmarks and head displacement, velocity, and acceleration along roll, pitch and yaw derived from AFAR toolbox [25]. In addition to the dynamics of the head, its orientation and changes along the same were also quantized as:

(1) Displacement, velocity and acceleration along roll, pitch and yaw
(2) Duration of head oriented left / right (similarly up / down and clockwise (CW) / anticlockwise (ACW))
(3) Mean duration of head aversion
(4) Rate of change of head orientation (left to right, up to down, and CW to ACW)

Facial landmarks were also used to extract eye-based activity such as distance between eyelids, duration of eye-closure and blinking rate. The following is an enumeration of features obtained using facial landmarks.

(1) Displacement, velocity and acceleration for 49 facial landmarks
(2) Distance between eyelids for each eye
(3) Eye-closure duration for each eye
(4) Blinking rate for each eye

All the features described above were summarized as mean, minimum, maximum, standard deviation, variance and interquartile range except the blinking rate and mean duration of head aversion. This resulted in 137 features that capture face and head activity. We followed the convention from [5] towards feature extraction.

*Speech Behavior.* The speech behavior features were extracted using manually segmented utterances from the recorded audio. Segmented utterances from audio were then used to identify inter-individual pauses. An inter-individual pause was defined as the time elapsed between end of the current utterance and the beginning of the next utterance. These utterances were characterised by speaker change.

Speech Behavior features include duration of turns, duration of overlapping speech, and duration of inter-individual pause. To resolve the ambiguity of the ownership of pauses and overlaps, the convention from Jaffe et al. [18] was adopted. Overlapping speech was assigned to the listener while the inter-individual pause was attributed to the speaker. All features were normalized for the duration of the interaction and quantified as sum (total duration), mean, minimum, maximum, standard deviation, variance and interquartile range. The resultant 21 features capturing the turn dynamics of the dyadic interaction were also used for depression prediction.

(1) Duration of inter-individual pauses and overlapping speech
(2) Spoken duration

*Linguistic.* The Linguistic Inquiry Word Count (LIWC) framework [27] was used to capture various attributes of the spoken language. About 93% of spoken words were found in the LIWC dictionary following which the coverage variable "Dic" was dropped because of its lack of relevance to the prediction problem. To exclude features that were primarily targeted at the written language, 9 punctuation categories (represented as AllPunc, Period, Comma, Colon, SemiC, Quote, Apostro, Parenth, OtherP in LIWC) were eliminated from the study. The 83 resultant features then spanned across several LIWC categories such as linguistic processes, psychological processes, cognitive processes, perceptual processes, biological processes, personal concerns and spoken assents and disfluencies (disfluencies such as repetition of partial words or phrases were represented through a hyphen in the transcription process).

*Prosody.* Prosody features consisted of the eGeMAPS [12] feature set from Opensmile (version 3.0) [13] framework. Manually segmented audio identified by the speaker was used to select the voiced portions of the children in the interaction and corresponding features were used. 27 low-level descriptors (LLDs) identified in [1] along with their first and second-order differences were of interest towards prosody. Briefly, the LLDs consisted of pitch, jitter, shimmer, loudness, Harmonic-to-Noise-Ratio (HNR), formant

frequencies, MFCCs, Pitch-to-Harmonics difference ratio, Hammarberg index and slope in different spectral bands. These features were then summarized using mean, minimum, maximum, standard deviation, variance and interquartile range for prediction. This resulting 486 features represented prosody and its dynamics.

Despite some recent evidence [31] that articulatory features outperform eGeMAPS, our choice of eGeMAPS is based on the finding that it is a robust predictor across cultures and languages [1].

## 3.3 Manual annotation by experts

Living In Family Environments (LIFE) system is an observational coding framework proposed by Hops et al. [17]. It includes separate codes for non-verbal and verbal measures of affective interpersonal behaviors displayed by family members. The non-verbal and verbal codes are combined to define four constructs–aggressive, dysphoric, positive and other (mostly neutral). The manually annotated constructs followed a stop-frame protocol where onsets were annotated on every behavior change. Two highly trained coders supervised directly by a master coder and indirectly by an investigator with significant expertise in the LIFE coding protocol [17] annotated the videos. Inter-rater reliability was established prior to the initiation of coding and monitored regularly, throughout. Inter-rater reliability (Krippendorf's alpha) was 0.66 [33]. The relevance of constructs to depression was demonstrated by [30] in which construct-coded mother's behavior during mother-child problem-solving interaction were found to be significant predictors for future onset of adolescent depression. The LIFE system has also been shown to discriminate between depressed and non-depressed adolescents on their own coded behavior. Such outcomes motivated the use of observer ratings of constructs as a baseline with which to evaluate computational multimodal measures.

Separate Support Vector Machines (SVM) were constructed for manual measures (referred to as "constructs" in the LIFE scheme) and multimodal computational measures. Each of the four LIFE constructs was assigned duration and frequency (analogous to the rate-per-minute variable used in [30]) attributes. The duration of a given construct is defined as the time elapsed between the onset of one construct and the onset of the next. Durations were summarized using 6 summary statistics: mean, minimum, maximum, variance, standard deviation, and inter-quartile range corresponding to that for the multimodal computational features. In addition, the frequency (number of instances of a given construct / duration of the interaction) of each construct was also included. This resulted in 28 features that represent manually coded constructs.

## 3.4 Experimental Design

The classification setup included Support Vector Machines (SVM) with linear and polynomial kernels. The choice of SVM was based on its generalization capabilities due to the max-margin criterion and previous use in detecting depression [5, 8]. In addition to the choice of the kernels, the hyperparameters also included the cost of misclassification *C*. All features were Z-score normalized. Multimodal classifiers through early and late fusion strategies described below were also used.

*3.4.1 Early fusion.* Early fusion of features from different modalities included concatenation of unimodal features before being used to train the classifier.

*3.4.2 Late fusion.* For late fusion, decisions from trained unimodal classifiers were combined using a weighted sum criterion. Unlike the early fusion strategy, the late fusion had additional hyperparameters i.e. the fusion weights used to combine the unimodal decisions. These weights were determined using a grid-search on a [0, 1] scale in steps of 0.1 and the best setting was identified using a nested cross-validation described below.

A five-fold testing with a nested four-fold cross-validation was used to determine the best hyperparameters in all experiments including the choice of optimal features, and fusion weights for late fusion wherever necessary. All hyperparameters were determined based on the mean accuracy over the validation sets in the nested step. These were then used to report the test performance over the unseen data in the outer loop. Prediction performance was evaluated using mean accuracy, positive (PA) and negative (NA) class agreements across the five outer folds. Definitions of PA and NA are described in Girard et al. [14].

## 3.5 SHAP for feature selection

Approximate Shapley values have been traditionally used for interpreting model predictions. They were used to identify top features that influenced classifier decision in the multimodal prediction of mania [3] and history of depression [5]. Their model-agnostic nature and the availability of fast solutions for the approximate computation of Shapley values through the SHAP framework are some of the factors that contributed to their pervasive use in affective computing. Shapley values (Equation 1) correspond to the contribution made by a given feature when it was present and marginalized. They can be approximated using the Kernelized SHAP [20] approach to determine the SHAP values.

$$\varphi_j(v) = \sum_{S \subseteq \{1,2,\ldots,m\} \setminus \{j\}} \frac{|S|!\ (m - |S| - 1)!}{m!} (v(S \cup \{j\}) - v(S)) \quad (1)$$

where $\varphi_j$ is the Shapley value for the $j^{th}$ feature, $v$ is the prediction model, $m$ is the total number of features and $S$ is a subset of features.

The strength of a feature's influence on a prediction is directly proportional to absolute SHAP value and polarity determines its inclination towards a given category. To determine feature ranking over a dataset, given a prediction model, the mean absolute SHAP value over the dataset calculated per feature is used to define how much they influence the prediction model. Higher the SHAP value, more discriminative 'power' is held by the feature. We propose features ranked by their mean absolute SHAP value be iteratively added in steps of k=10 as long as they demonstrate an increase in the mean validation accuracy in the nested folds (see Algorithm 1). Differences in the optimal features between outer folds were observed. For example, say, $OF_1$ were the optimal features for fold-1 while $OF_2$ were the optimal features for fold-2 where $OF_1 \neq OF_2$. In such case, to prevent excessive feature rejection, the union of optimal features ($OF_1 \cup OF_2$) across outer folds was used for testing. Figure 1 presents an increasing validation accuracy trend

**Table 1: Detection performance for children of mothers with history of depression using child features. Bold indicates the best performance.**

| | No feature selection | | | | SHAP-based feature selection | | | |
|---|---|---|---|---|---|---|---|---|
| Modality | #features | Acc | PA | NA | #features | Acc | PA | NA |
| Observer ratings | 28 | 0.595 | 0.485 | 0.667 | N/A | N/A | N/A | N/A |
| Action Units | 156 | 0.602 | 0.596 | 0.559 | 130 | 0.606 | 0.601 | 0.565 |
| Face & Head dynamics | 137 | 0.522 | 0.425 | 0.408 | 120 | 0.499 | 0.426 | 0.371 |
| Speech Behavior | 21 | 0.559 | 0.542 | 0.565 | 21 | 0.559 | 0.542 | 0.565 |
| Linguistic | 83 | 0.591 | 0.532 | 0.615 | 78 | 0.564 | 0.502 | 0.580 |
| Prosody | 486 | 0.645 | 0.588 | 0.646 | 197 | 0.740 | 0.716 | 0.729 |
| Early Fusion | 883 | 0.634 | 0.603 | 0.638 | 546 | 0.714 | 0.687 | 0.703 |
| Late Fusion | 883 | 0.593 | 0.539 | 0.602 | 546 | 0.680 | 0.649 | 0.645 |
| Shapley on Early Fusion | N/A | N/A | N/A | N/A | **225** | **0.769** | **0.752** | **0.759** |

in different folds as a result of the proposed feature selection using SHAP.

**Algorithm 1** SHAP-based greedy feature selection for $j^{th}$ outer fold in nested cross-validation scenario

**Require:** Dataset $D$ with $m$ features, where $D^m_{v_{ij}}$ is $i^{th}$ validation set nested in the $j^{th}$ outer fold. *meanSHAP* function returns the mean SHAP values from validation sets in $j^{th}$ outer fold.

**Ensure:**

```
Optimal features OF_j ← list()
BestValidationAccuracy ← 0.5
k ← 10        ▷ Let features be evaluated in increments of k = 10
rankedFeatures ← meanSHAP(D^m_{v_j})
for i in range(0, m, k) do
    featuresStudied ← rankedFeatures[i : (i + 1) * k]
    featuresToEval ← OF_j + featuresStudied
    ValidationAccuracy ← trainSVMClassifier(featuresToEval)
    if ValidationAccuracy > BestValidationAccuracy then
        OF_j ← featuresToEval
        BestValidationAccuracy ← ValidationAccuracy
    end if
end for
return OF_j
```

## 4 RESULTS

Detection performance for children of depressed mothers is summarized in Table 1. With no feature selection, both unimodal and multimodal features can reliably detect the same. However, there is a clear distinction in the efficacy of various modalities in achieving it. The prosody features achieved 64.5% accuracy and is the best performing modality including against the fusion methods such as early and late fusions. Early fusion of modalities with 883 features achieved comparable performance with prosody, however, at a cost of 1.8× more features. Late fusion on the other hand underperformed as compared to early fusion. Among unimodal features, action units and linguistic features achieved 60.2% and 59.1% accuracy respectively. This is despite almost 2× fewer features in linguistic modality. Speech behavior with a relatively small set of features, can predict at 55.9% accuracy. However, the face and head dynamics performs barely above chance-level with 137 features. When compared against the manually annotated constructs, most unimodal features (except face and head dynamics and speech behavior) outperform them. Fusion methods, particularly early fusion exceeds the constructs by over 6% accuracy while late fusion performs comparably to constructs. Detection using prosody also outperforms constructs by 5% accuracy. Despite the step size k=10 used in feature selection, the final features are non-multiples of k because of the union of the optimal features across outer folds. However, the optimal features observed per-fold were multiples of k.

The SHAP-based feature selection has offered performance improvements for prosody and fusion methods. Prosody features experienced a close to 10% improvement in accuracy with 2.5× fewer features. Interestingly, such behavior was not observed with any other modality. This suggests that the SHAP-based feature selection is likely sensitive to the nature of the modality. Barring face and head dynamics and linguistic modalities where the performance dropped, other modalities have performed similarly to no feature selection with a limited reduction in the number of features. On comparing the class-level agreements (i.e. PA and NA), we notice that positive agreement increased by about 13% while negative agreement increased by 8.3% for prosody. Late fusion performed similar to prosody, where accuracy increased by 9% with PA and NA increased by 11% and 4.3% respectively. This pattern among feature selection based on ranking through SHAP values suggests that it prioritizes preserving features that correspond to the positive (i.e. children of depressed mothers) class. While early fusion has experienced an 8% increase in accuracy, it has offered at-par improvements to both positive and negative classes.

To understand how well the feature selection strategy works in the multimodal context, we further performed SHAP-based feature selection on the early fusion of previously selected unimodal features (see *Shapley on Early Fusion* in Table 1). Results show that this considerably reduces the number of features by about 2.4× with a 5.5% improvement in accuracy. This feature selection for multimodal features approach outperforms the constructs by over 17%.

Having established that differences in child behavior can be reliably detected, further attempts were made to identify the features that differ the most between them. Table 2 presents the 20

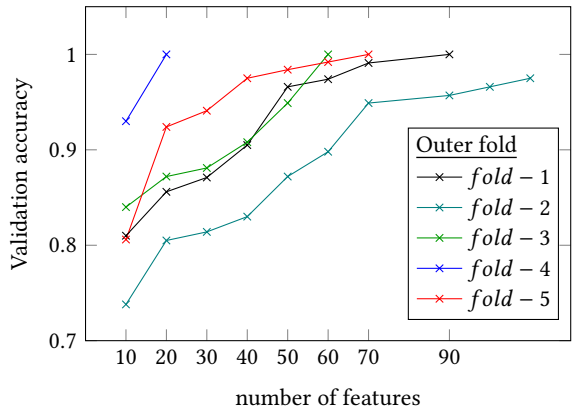


**Figure 1: Effect of feature selection on validation accuracy in Early fusion with SHAP-selected unimodal features.**

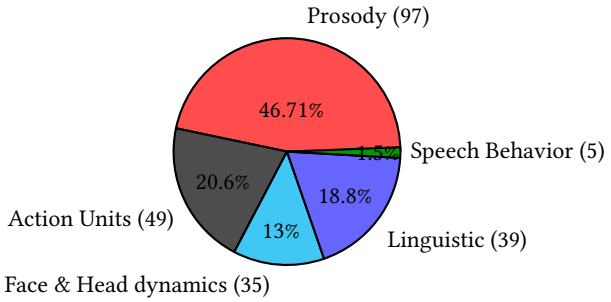


**Figure 2: Relative contribution of each modality and the number of features each includes. The total number of features for all modalities sums to 225.**

most contributing features from the best performing *Shapley on Early Fusion* classifier in Table 1. While SHAP values provide an understanding of individual features, to accommodate for the large number of features in each modality, modality-wise contributions were calculated as the ratio between mean absolute SHAP values of all features from a given modality to mean absolute SHAP values of all features across modalities (Equation 2).

$$C_{modality_m} = \frac{\sum_j \sum_i |SHAP_{ijm}|}{\sum_k \sum_j \sum_i |SHAP_{ijk}|} \quad (2)$$

where $C_{modality_m}$ is the normalized contribution of modality $m$ and $SHAP_{ijk}$ corresponds to the SHAP value for the $i^{th}$ feature from the $k^{th}$ modality in the $j^{th}$ sample.

Modality contributions through Figure 2 shows that the prosody constitutes the most discernable features. With 97 (5× reduction wrt original unimodal prosody without feature selection) features out of the final 225 features (i.e. 43% of features), its modality contribution is 46.1%. This is followed by action units, where 21.8% features contributed 20.6% based on the SHAP values, a total 2.8× fewer features when compared to the original features. Similarly, the linguistic features with 17.3% (2.4× reduction) of the final features it contributed 18.8% and face and head dynamics with 15.5% (3.9× reduction) of the final features contributed 13%. Speech behavior had a 4.2× reduction in the number of features against its original 21 features with a marginal 1.5% modality contribution.

**Table 2: SHAP ranking of top-20 features grouped by modality. Modalities and features within modalities are listed in order of their relative contribution. *Depression* column shows the overall ranking of each feature across all modalities and the sign of correlation between feature with predicted depression. Red indicates positive correlation; purple, negative correlation.**

| Modality | Feature | Depression |
|---|---|---|
| **Prosody** | MFCC4 rate of change | 3 |
| | spectral slope 0-500 change | 4 |
| | loudness change | 5 |
| | F1AmpLogRelF0 rate of change | 6 |
| | spectral flux change | 8 |
| | alpha ratio | 11 |
| | F1 bandwidth change | 16 |
| | F2AmpLogRelF0 rate of change | 17 |
| | F1 rate of change | 20 |
| **Action Units** | AU 4 mean occurrence | 2 |
| | AU 4 occurrence variance | 7 |
| | AU 4 occurrence IQR | 14 |
| | AU 15 occurrence max | 15 |
| **Linguistic** | Body | 1 |
| | Differentiation | 9 |
| | Home | 10 |
| | Focus present | 12 |
| | Insight | 13 |
| | six-letter word | 19 |
| **Face and Head dynamics** | landmark velocity IQR | 18 |

## 5 DISCUSSION

**Computational measures outperform manual annotations:** Comparing manually annotated observer ratings and computational measures through multimodal features, multimodal features were found to be more informative of the impact of depression on child behavior. In addition to being better predictors (due to higher accuracy), they also provided insight into how different low-level interpretable features differ between the groups of children. Manual annotations on the other hand are limited by their constituent high-level descriptions of behavior and how reliably they can be identified by annotators. Despite the advantages of using computational multimodal features over manually coded behavior, they are susceptible to the limitations of the prediction framework used for feature extraction and are relatively more sensitive to the data collection setup. For example, the face and head dynamics features extracted from AFAR toolbox [25] are limited at extreme head pose and occlusions due to the failure of face detection and tracking algorithms. Similarly, data collection requirements such as near-frontal camera orientation and speaker-separated audio in proximal environments make data collection for multimodal features challenging.

A novel feature selection approach through a model interpretation solution (Shapley Additive Explanations) demonstrated its efficacy with prosody features. Unlike other modalities, feature selection in prosody resulted in 2.5× fewer features and a 10%

increase in accuracy. This shows the sensitivity of the feature selection approach with modality. The effectiveness of prosody features in predicting behavioral differences is not new. Yang et al. [34] showed that naive listeners can identify depression severity from vocal prosody. Our study furthers this understanding in the novel context of differences in child behavior due to mother's history of depression. While simple concatenation (early fusion) of unimodal features after feature selection did not result in any performance improvement, further feature selection to account for inter-modal interactions through the *Shapley on early fusion* resulted in the best performing model with 76.9% accuracy through reduced overfitting and 4× reduction in the original features.

***Body* and *Home*:** The top-20 features ranked by their SHAP values (Table 2) from the best performing *Shapley on Early fusion* model are a mixture of linguistic, prosody, action units and face and head dynamics modalities. The linguistic features captured physical processes related to body, differentiation of thoughts, words related to home environment, focus present, insight words and usage of complex language (six-letter words). Body related words in the dataset include references to hair, sweaty, fat, shoe, tongue, hands etc. In the context of mother-child interaction, their usage was found to be associated with personal hygiene concerns in children. To list a few, these include: brushing hair, choice of dreads, taking shower and ear piercings. Interestingly such references were lower in the children of depressed mothers when compared to the children of non-depressed mothers. Differentiation words such as *but*, *except* and *versus* were seldom used by children of depressed mothers. Words related to the home environment through references to room, shower, bed, chores, neighbors, landlord, rent, staying home alone, using the bathroom all the time, issues with commute such as driving home and back to school, organising stuff around the house (ex. *don't put that in the hall closet*) were found to be higher in the depressed group. Lower references to personal hygiene but increased references to home environment suggest that children of mothers with history of depression were invested in discussing issues related to life at home and/or school over concerns about personal hygiene that are commonly observed among children. Similarly focus on present and usage of words of insight such as *think, know* and *consider* that capture thinking styles through cognitive mechanisms, and complex language was found to be higher among children of depressed mothers.

**Effectiveness of prosody dynamics:** We found that the dynamics of prosody through their first and second-order differences were highly influential in detecting the children of depressed mothers over the original eGeMAPS prosody features. For example, the rate of change of MFCC4 component was lower in children with a depressed mother. Similar trend was observed with change in spectral slope at 0-500Hz. The MFCC and spectral slope features capture energy in different frequency bands during speech production [8]. Ozdas et al.'s work [26] on jitter and spectral slope for predicting suicide and depression suggests that speech in near-term suicidal and depressed individuals has increased energy in the high frequency (300–3000 Hz) bands. As symptoms subside, the energy activity returns to the lower band [29]. While our findings on spectral slope do not directly compare the low vs. high-frequency bands, they suggest that changes in lower band are limited among the children of depressed mothers. Alpha ratio, the ratio of energy between 50Hz to 1kHz and 1kHz to 5kHz suggests an increased activity in lower frequency bands among children of depressed mothers. This is contrary to our earlier [26, 29] understanding of spectral features where high-frequency activity was observed with depression and suicidal ideation. Spectral flux captures the differences in energy between successive speech frames. This is also closely related to the perceived vocal loudness. Both features capture vocal arousal [12]. On these lines, we found that both spectral flux and loudness exhibit similar behavior that show that changes in arousal (due to change in loudness) and its rate of change (change in spectral flux) are higher among children of depressed mothers. This suggests that conflict resolution involving mother with a history of depression leads to more dynamic conversations than their non-depressed counterparts.

Formant frequencies are associated with changes in the vocal-tract resonance cavity and represent the articulatory effort in speech production. We notice that both the changes in the first formant frequency (F1) bandwidth and the rate of change of its (F1) frequency is higher in children of depressed mothers over children of non-depressed mothers. Though literature [8] shows some evidence of the role of formant frequencies in depression, limitations in accurately capturing the relationship between articulatory behavior and formants, and improper glottal closure resulted in lack of replicable results. These factors make it hard to make conclusive inferences. Relative energy in first (F1) and second formants (F2) with respect to the fundamental frequency (F0) correspond to the perceived voice quality [22]. They capture excessive nasal and clicky sounds. We found that the rate of change of voice quality in the first (shown as F1AmplogRelF0 in Table 2) and second formants (shown as F2AmplogRelF0) also differ among children based on their mother's history of depression.

**Negative affect and arousal:** Both AU 4 (lowering eyebrows) and AU 15 (lip corner depressor) associated with negative emotions [4] appear in the top contributing features. AU 4 is found in expressions of anger and fear while both AUs 4 and 15 are observed in sadness. AU 4 occurrence appears in top features as various summary statistics. It suggests increased expressions of negative emotion occurrences along with high variance and inter-quartile range i.e. more dynamic expressions were observed in children whose mother has a history of depression. AU 15 occurrence, particularly sadness was also highly likely among children of depressed mothers. Face dynamics had only one top feature–arousal measured through landmark velocity was found to be flat (or less dynamic) in the children with depressed mothers.

**Limitations:** Four limitations may be noted. One, the context of mother-child interaction raises the question whether the behavioral differences among children we found would generalize to their interactions with other family members, peers, or other persons. Two, participants were low-income, predominantly white, and from a single city in the US. Whether the findings would apply to families in other parts of the US or other countries or cultures remains to be found. Three, the choice of classifier leaves room for exploring ensemble approaches such as random forest and XGBoost. Four, generalization of the proposed feature selection approach across different prediction frameworks remains as future work.

## 6 CONCLUSION

Our work addresses the question of behavioral differences in adolescent children of mothers with and without depression. We frame the problem as a detection task defined by mothers' depression status and use multimodal computational measures of child behavior. Using a novel SHAP-based feature selection approach, we found that reliable prediction can be performed using both unimodal (especially, prosody) and multimodal features. Prosody was the single most informative feature set. That was followed by facial action units, linguistic features, face and head dynamics, and speech behavior. A comparison model that used subjective manually annotated features paled in comparison to that from the leading computational models.

SHAP-based feature selection significantly reduced the number of features and increased model performance. Shapley analysis revealed that children of depressed mothers have different spectral dynamics, voice quality and amplitude. Their facial actions reveal reduced arousal and greater negative affect. Linguistic features differ as well. Such naunced behavioral differences between children with and without depressed mothers extend the existing understanding [15] of the influence of mother's depression on their offspring.

The study's limitations included: Dyadic interactions were limited to mothers and children; diversity was limited to low-income participants from a single city and country; and alternative classifiers and frameworks remain to be explored.

## ACKNOWLEDGMENTS

This research was supported in part by the U.S. National Institutes of Health through U.S. National Institute of Mental Health award MH096951.

## REFERENCES

[1] Sharifa Alghowinem, Roland Goecke, Julien Epps, Michael Wagner, and Jeffrey F Cohn. 2016. Cross-cultural depression recognition from vocal biomarkers.. In *Interspeech*. 1943–1947.

[2] Sharifa Mohammed Alghowinem, Tom Gedeon, Roland Goecke, Jeffrey Cohn, and Gordon Parker. 2020. Interpretation of depression detection models via feature selection methods. *IEEE Transactions on Affective Computing* (2020).

[3] Pınar Baki, Heysem Kaya, Elvan Çiftçi, Hüseyin Güleç, and Albert Ali Salah. 2022. A Multimodal Approach for Mania Level Prediction in Bipolar Disorder. *IEEE Transactions on Affective Computing* 13, 4 (2022), 2119–2131.

[4] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest* 20, 1 (2019), 1–68.

[5] Maneesh Bilalpur, Saurabh Hinduja, Laura A Cariola, Lisa B Sheeber, Nick Allen, László A Jeni, Louis-Philippe Morency, and Jeffrey F Cohn. 2023. Multimodal Feature Selection for Detecting Mothers' Depression in Dyadic Interactions with their Adolescent Offspring. (2023).

[6] Laura A Cariola, Saurabh Hinduja, Maneesh Bilalpur, Lisa B Sheeber, Nicholas Allen, Louis-Philippe Morency, and Jeffrey F Cohn. 2022. Language Use in Mother-Adolescent Dyadic Interaction: Preliminary Results. In *ACII*. IEEE.

[7] James C Coyne. 1976. Depression and the response of others. *Journal of abnormal psychology* 85, 2 (1976), 186.

[8] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech communication* 71 (2015), 10–49.

[9] Paul Ekman, Wallace V Friesen, and Joseph C Hager. 2002. *Facial Action Coding System: Facial action coding system: the manual: on CD-ROM*. Research Nexus.

[10] Karen A Ertel, Janet W Rich-Edwards, and Karestan C Koenen. 2011. Maternal depression in the United States: Nationally representative rates and risks. *Journal of women's health* 20, 11 (2011), 1609–1617.

[11] Itir Onal Ertugrul, Jeffrey F Cohn, László A Jeni, Zheng Zhang, Lijun Yin, and Qiang Ji. 2020. Crossing domains for au coding: Perspectives, approaches, and measures. *IEEE transactions on biometrics, behavior, and identity science* 2, 2 (2020), 158–171.

[12] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 7, 2 (2015), 190–202.

[13] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *ACM Multimedia*. 1459–1462.

[14] Jeffrey M Girard, Wen-Sheng Chu, László A Jeni, and Jeffrey F Cohn. 2017. Sayette group formation task (GFT) spontaneous facial expression database. In *FG*. IEEE, 581–588.

[15] Sherryl H Goodman, Hannah FM Simon, Amanda L Shamblaw, and Christine Youngwon Kim. 2020. Parenting as a mediator of associations between depression in mothers and children's functioning: A systematic review and meta-analysis. *Clinical Child and Family Psychology Review* 23 (2020), 427–460.

[16] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. *The distress analysis interview corpus of human and computer interviews*. Technical Report. UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES.

[17] Hyman Hops, Betsy Davis, and Nancy Longoria. 1995. Methodological issues in direct observation: Illustrations with the Living in Familial Environments (LIFE) coding system. *Journal of Clinical Child Psychology* 24, 2 (1995), 193–203.

[18] Joseph Jaffe, Beatrice Beebe, Stanley Feldstein, Cynthia L Crown, Michael D Jasnow, Philippe Rochat, and Daniel N Stern. 2001. Rhythms of dialogue in infancy: Coordinated timing in development. *Monographs of the society for research in child development* (2001), i–149.

[19] Anis Kacem, Zakia Hammal, Mohamed Daoudi, and Jeffrey Cohn. 2018. Detecting depression severity by interpretable representations of motion dynamics. In *FG*. IEEE, 739–745.

[20] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *NeurIPS* 30 (2017).

[21] Wilson E Marcílio and Danilo M Eler. 2020. From explanations to feature selection: assessing SHAP values as feature selection mechanism. In *2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI)*. Ieee, 340–347.

[22] Shahan Ali Memon. 2020. Acoustic Correlates of the Voice Qualifiers: A Survey. *arXiv preprint arXiv:2010.15869* (2020).

[23] Benjamin W Nelson, Lisa Sheeber, Jennifer Pfeifer, and Nicholas B Allen. 2021. Psychobiological markers of allostatic load in depressed and nondepressed mothers and their adolescent offspring. *Journal of Child Psychology and Psychiatry* 62, 2 (2021), 199–211.

[24] Benjamin W Nelson, Lisa Sheeber, Jennifer H Pfeifer, and Nicholas B Allen. 2021. Affective and Autonomic Reactivity During Parent–Child Interactions in Depressed and Non-Depressed Mothers and Their Adolescent Offspring. *Research on Child and Adolescent Psychopathology* 49, 11 (2021), 1513–1526.

[25] Itir Onal Ertugrul, László A Jeni, Wanqiao Ding, and Jeffrey F Cohn. 2019. AFAR: A Deep Learning Based Tool for Automated Facial Affect Recognition. In *FG*. IEEE.

[26] Asli Ozdas, Richard G Shiavi, Stephen E Silverman, Marilyn K Silverman, and D Mitchell Wilkes. 2004. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *IEEE transactions on Biomedical engineering* 51, 9 (2004), 1530–1540.

[27] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.

[28] Anupama Ray, Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee, and Ritu Garg. 2019. Multi-level attention network using text, audio and video for depression prediction. In *9th AVEC challenge*. 81–88.

[29] Klaus R Scherer. 1979. Nonlinguistic vocal indicators of emotion and psychopathology. *Emotions in personality and psychopathology* (1979), 493–529.

[30] Orli S Schwartz, Paul Dudgeon, Lisa B Sheeber, Marie BH Yap, Julian G Simmons, and Nicholas B Allen. 2012. Parental behaviors during family interactions predict changes in depression and anxiety symptoms during adolescence. *Journal of abnormal child psychology* 40, 1 (2012), 59–71.

[31] Nadee Seneviratne and Carol Espy-Wilson. 2022. Multimodal depression classification using articulatory coordination features and hierarchical attention based text embeddings. In *ICASSP*. IEEE, 6252–6256.

[32] Jarne Verhaeghe, Jeroen Van Der Donckt, Femke Ongenae, and Sofie Van Hoecke. 2022. Powershap: A power-full shapley feature selection method. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 71–87.

[33] Torsten Wörtwein, Lisa B Sheeber, Nicholas Allen, Jeffrey F Cohn, and Louis-Philippe Morency. 2021. Human-Guided Modality Informativeness for Affective States. In *ACM ICMI*. 728–734.

[34] Ying Yang, Catherine Fairbairn, and Jeffrey F Cohn. 2012. Detecting depression severity from vocal prosody. *IEEE transactions on affective computing* 4, 2 (2012), 142–150.