Multimodal Feature Selection for Detecting Mothers' Depression in Dyadic Interactions with their Adolescent Offspring

Maneesh Bilalpur[1], Saurabh Hinduja[2], Laura A. Cariola[3], Lisa B. Sheeber[4], Nick Allen[5], László A. Jeni[6], Louis-Philippe Morency[7] and Jeffrey F. Cohn[1,2,6]

[1] Intelligent Systems Program, University of Pittsburgh, Pittsburgh, USA
[2] Department of Psychology, University of Pittsburgh, Pittsburgh, USA
[3] Clinical and Health Psychology, University of Edinburgh, Edinburgh, UK
[4] Oregon Research Institute, Oregon, USA
[5] Department of Psychology, University of Oregon, USA
[6] Robotics Institute, Carnegie Mellon University, USA
[7] Language Technology Institute, Carnegie Mellon University, USA

Abstract— **Depression is the most common psychological disorder, a leading cause of disability world-wide, and a major contributor to inter-generational transmission of psychopathology within families. To contribute to our understanding of depression within families and to inform modality selection and feature reduction, it is critical to identify interpretable features in developmentally appropriate contexts. Mothers with and without depression were studied. Depression was defined as history of treatment for depression and elevations in current or recent symptoms. We explored two multimodal feature selection strategies in dyadic interaction tasks of mothers with their adolescent children for depression detection. Modalities included face and head dynamics, facial action units, speech-related behavior, and verbal features. The initial feature space was vast and inter-correlated (collinear). To reduce dimensionality and gain insight into the relative contribution of each modality and feature, we explored feature selection strategies using Variance Inflation Factor (VIF) and Shapley values. On an average collinearity correction through VIF resulted in about 4 times feature reduction across unimodal and multimodal features. Collinearity correction was also found to be an optimal intermediate step prior to Shapley analysis. Shapley feature selection following VIF yielded best performance. The top 15 features obtained through Shapley achieved 78% accuracy. The most informative features came from all four modalities sampled, which supports the importance of multimodal feature selection.**

I. INTRODUCTION

Approximately one in seven adolescents between 10 and 19 years of age experience a mental disorder [26]. Depression is especially common. Past studies[3], [27] have found that the well-being of parents plays an important role in the well-being of their children. Children with parents who have recurrent episodes of depression are at significantly increased risk for depression and other disorders. Findings such as these suggest that family environment profoundly influences adolescent mental well-being.

Computational approaches to depression assessment have been a popular problem in the affective computing community [7], [36], [35]. Using objective measures of behavior to detect depression, they seek to overcome some of the limitations of self-report based ratings [5]. Most of this work has focused on a small number of constrained settings, such as clinical interviews [14], [15], interactions with a virtual agent [17], and reading aloud [36], [35]. Ours may be the first computational study to consider the influence of depression between family members, specifically, mothers and their adolescent offspring. To contribute to developmental behavioral and clinical science, we seek to understand how depression is communicated within families. Depression was defined as history of treatment for depression together with current or recent symptoms.

Due to the sheer number of features involved in computational approaches to depression detection and the limited data typically available, models are at risk for overfitting. To address this problem, dimensionality reduction methods such as Principal Component Analysis (PCA) have been used. But PCA often leads to difficulty in interpreting the contribution of features. Dropping highly correlated features is another alternative. Because that solution is limited to linear pairwise relations, collinearity among multiple features fails to be considered. To overcome these limitations, we use Variance Inflation Factor (VIF), a feature-centric approach to determine collinearity.

Features were multimodal and included a wider range than in most previous studies. They encompassed head orientation and dynamics, facial action units and gaze, speech behavior, and speech. To afford interpretability and because the number of participants was modest (n = 152), a priori features rather than learned ones were used.

Given the very large number of multimodal features, collinearity within and between modalities was common. We pursued two approaches in combination for collinearity correction and feature reduction: Variance Inflation Factor (VIF) and Shapley analysis[33], [21]. Together, they achieved interpretability and improvement in performance. We sum-

Research reported in this publication was supported in part by the US National Institutes of Health under Award Number MH096951. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

978-8-3503-4544-5/23/$31.00 ©2023 IEEE

marize our contributions as follows:

1) Investigate the influence of depression in social interactions between mothers and their adolescent offspring. Previous studies of social interaction in depression are limited to solitary tasks or structured interviews between unrelated individuals.
2) Depression was defined as history of treatment for depression together with elevations in current or recent symptoms, which is relevant to developmental outcomes in children.
3) Unlike most previous works (see related works), we use both verbal and a holistic set of non-verbal behavior for depression detection and evaluate how they influence the prediction.
4) Explore the efficacy of a variety of unimodal and multimodal features in depression detection.
5) Demonstrate that many features for depression detection are highly collinear and that by reducing collinearity through VIF followed by Shapley analysis substantial improvement in detection accuracy can be achieved.

II. Related Works

Previous work in depression detection has focused on clinical interviews and other constrained tasks (e.g., reading digits). In audio-video recordings of clinical interviews, Dibeklioglu et al.[9] used face and head dynamics and prosody features for depression detection. They performed dimensionality reduction of landmarks and head pose (roll, pitch and yaw) parameters using autoencoders. These compact features were then used to calculate displacement, velocity and acceleration of landmarks and head pose changes. Frame-level dynamic features were then consolidated using improved Fisher-vector representation to obtain video-level features. After feature selection using the Min-Redundancy Max-Relevance algorithm, best performing model achieved a mean accuracy of 78% across different levels of depression severity.

The utility of action units towards depression prediction was demonstrated by Girard et al.[14], [15]. They performed comparisons of action unit base rates among patients treated for depression. Manual FACS coding was done for a small portion of the data which was then used as ground truth to train a classifier to obtain AU predictions for entire data. It was found that among depressed subjects, the action units corresponding to affiliation (AU12 and AU15) occurred seldom while those associated with isolation (AU10 and AU14) were found more common.

Facial landmarks, spectrograms of audio and word embeddings from transcription were used in Haque et al.[18] to extract sentence-level multimodal embeddings. Unlike studies where video-level features were consolidated through clustering based approaches or summary statistics, they performed sentence-level depression prediction. The sentence-level multimodal embeddings were used to predict (both classification and regression) for depression diagnosis and depression severity as measured using a self-reported PHQ-8 questionnaire.

Morales et al.[22] proposed a syntax-informed fusion for depression prediction using audio, syntax of spoken words and action units. Instead of using multimodal features as independent features, they extracted syntax-conditioned features such as audio features in the presence of verb usage. It was found that syntax-informed features improve performance and highlight novel features that were thought to be less informative in an early fusion setting.

In works related to hybrid classifiers, Yang et al.[37] used spoken portions of subject-agent interaction to predict depression. Unlike most works mentioned, where depression was solely evaluated from verbal and/or non-verbal cues, this work used the self-report on personality, mood and other physical and mental conditions to train a text based SVM prediction model. They also employed a piecewise training strategy of prediction models involved. They trained a CNN model for audio and visual features independently for PHQ-8 score prediction followed by a fully connected model and then fused the resultant score prediction with the predicted physical and mental conditions of the subject for depression diagnosis.

Generalization of features across datasets 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 three different datasets for depression prediction. 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 (clinical interview) and AVEC (interaction with a computer) datasets found generalization on the Pitt dataset (clinical interview) challenging. They found that a subset of features from gaze and prosody generalized across datasets.

These studies vary in how they defined depression. Some used diagnostic criteria and others used symptom severity as ascertained by self report measures or clinical interviews. We defined depression as history of prior treatment together with current or recent symptoms. From a developmental perspective, mothers' history of depression is what matters in influencing child outcomes [20]. As noted above, previous studies were limited to depression in solitary or structured social contexts like interview between unrelated persons. Our work focuses on related individuals, specifically family members in social interactions.

III. Dataset and Methods

In this section we describe participants, observational procedures, and feature extraction.

A. Participants

Participants were 180 low-income women and their adolescent children, aged 11–14. Two groups of women were recruited: a depressed group, selected for a history of treatment for depression together with elevated depressive symptoms at the time of recruitment (PHQ-9 cut-off score

> 10; mean = 12.32, SD = 5.84) and a non-depressed group, selected for no history of treatment for depression, no or low levels of current depressive symptomatology (PHQ-9 cut-off score < 8; mean = 2.57, SD = 2.70), and no current (i.e., past month) mental health treatment for any mental health disorder. Exclusion criteria for participants of both groups included psychosis, other illness, or cognitive impairment that would interfere with participation (e.g., substance use that would render abstinence for the assessment difficult to tolerate). The Structured Clinical Interview non-patient version[12] was administered to confirm Depressed vs. Non-Depressed status.

Of the 180 participants, 152 participated in a lab-based dyadic interaction task that is the focus of this study. The depressed group consisted of 75 dyads (i.e., mother-child pairs). The remaining 77 dyads formed the non-depressed group. For further details on various assessment procedures, please see [23], [24]. Metadata and de-identified multimodal features will be made available to qualified researchers pending IRB approval.

B. *Observational procedures*

The mother-adolescent dyads from both groups participated in a Problem Solving Interaction task (PSI). The topic of conversation in the PSI was chosen to induce conflict between mother and child as per the Issues Checklist[31]. Audio was recorded at 16kHz with a dedicated microphone for each subject. Video was recorded at 30fps and 720p resolution using a dedicated camera for mother and child. Each interaction was designed to last 15 minutes. A 10-minute portion was manually segmented and transcribed for each interaction. For brevity, we use duration of the interaction when we refer to duration of the transcribed portion of the interaction throughout this work.

C. *Features for depression prediction*

Informed by prior research as reviewed above in Related Works, we extracted features for detection of depression. To enable interpretation of features, we use the summary statistics of various features discussed below.

1) Face and Head Dynamics (FHD): The orientation of mother‘s head, and facial expressions were captured through the angular orientation (roll, pitch and yaw) of head and facial landmarks extracted using AFAR toolbox[11], [25]. These were then used to define the dynamics of head and landmarks as detailed below.

The head angular orientation from AFAR was used to define head dynamics consisting of velocity and acceleration. We use the definitions from Dibeklioulgu et al.[9] to calculate the first derivative (difference between successive frames) of displacement as velocity and second derivative of displacement as acceleration. Dynamics along roll, pitch and yaw were extracted independently.

The angular orientation was used to extend features to include head oriented left/right, up/down, clockwise (CW) / anticlockwise (ACW) similar to Alghowinem et al.[2]. We define head oriented left/right as the subject facing the camera beyond one standard deviation (std) about the mean in yaw direction. Similarly for up/down we use pitch and for CW/ACW we use roll. These were then used to define head aversion. We quantize the head orientations using duration of head oriented up/down (similarly left/right, CW/ACW), mean duration of head aversion, rate of head orientation change.

The mean landmark dynamics (velocity and acceleration) were calculated using the 49 facial landmarks. In addition to these dynamics, for each eye, the distance between eyelids[4], eye-closure duration and blinking rate were calculated. We define blinking as an instance where the vertical displacement between eyelids is less than the mean minus one std. Eye-closure duration is defined as the duration for which the vertical displacement between eyelids is less than the mean minus one std. Blinking rate is defined as the number of instances of blinking per duration of the interaction.

The mean and standard deviation used to define thresholds were calculated using the entire dataset. All face and head orientation based dynamics were represented as the mean, minimum, maximum, standard deviation, variance, inter-quartile range of features described above. All duration and count based features were normalized with the duration of the interaction. This resulted in 143 features capturing summary statistics of various attributes of head and face.

2) Action Units (AU): The action unit predictions from the AU detection module of AFAR were used to extract the following features:

(a) Intensity for AUs 6, 10, 12, 14, 6+12 (positive affect or PA)
(b) Likelihood and duration of occurrence for AUs 1, 2, 4, 7, 15, 17, 23, 24

The intensity predictions are ordinal on a 0 to 5 scale while the occurrence predictions are probabilities indicating the likelihood of presence of a given AU per frame. Positive affect (PA) was defined as an average of intensities of AUs 6 and 12. To quantify duration, we threshold probabilities at 0.5 to define AU presence. Duration is defined as length of continuous AU presence normalized by the duration of the interaction. We did not use the duration for AUs 6, 10, 12, 14, 6+12 as including them with the rest of the features led to extreme multicollinearity (mean VIF>>100 across AU features). Note that AFAR does not provide intensity predictions for AUs 1, 2, 4, 7, 15, 17, 23 and 24.

The intensity, likelihood and duration features were consolidated using mean, minimum, maximum, standard deviation, variance, inter-quartile range. For actions units with intensity and occurrence predictions, we rely on intensity over occurrence to avoid inconsistencies between the same. 126 features summarizing intensity and occurrence dimensions of various facial muscle movements were then used in this study.

3) Speech Behavior (SB): The SB features were extracted using manually segmented utterances from the recorded audio. An utterance was defined as continuous speech activity with no more than 300ms of silence. The utterances segmented 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 beginning of the next utterance. These utterances were characterised by a speaker change.

SB features include duration of utterance by mother and child, duration of overlapping speech, and duration of inter-individual pause. All features were normalized for the duration of the interaction and quantified as sum (total duration), mean, minimum, maximum, std, variance and inter-quartile range. In addition to these features, we also calculated the rate of overlapping speech as the number of instances of speech overlap per duration of interaction. The resultant 29 features capturing the dialogue dynamics of the dyadic interaction were also used for depression prediction.

4) Verbal features (LIWC): The LIWC[30], [34] framework was used to extract 92 features from the verbal content of each interaction. On average, about 93% of words used in each interaction were present in the LIWC framework and analyzed. We drop the coverage (referred to as “Dic” in LIWC) variable and normalize (with duration of the interaction) word-count variable for our analysis. We only use the verbal content from mothers to extract the LIWC features.

The resulting 390 multimodal features across face and head dynamics (143), action units (126), speech behavior (29) and verbal features (92) of the mothers were used for depression prediction. Using the features corresponding to the children is of interest to future work.

D. Multicollinearity and Variance Inflation Factor

Multicollinearity (or collinearity) can be described as a linear relationship between variables. In regression analysis, using collinear features can lead to overfitting where model could fit the predictor variable with a high explained variance(R^2) despite no significant effect from any of the features[28] as a result of high variance in the parameter estimates. Variance Inflation Factor (VIF) can be used to eliminate highly collinear features while preserving the necessary features in their original form without the need for predictor information.

Variance Inflation Factor: Variance Inflation Factor (VIF) explains the contribution of each feature based on regression between features. The unexplained variance of a regression model for a given feature using the rest of the features as independent variables indicates how much “new information” the feature contains. Higher the unexplained variance of a feature, less collinear it is against the rest of the features. It helps in reducing feature redundancy by preserving features whose VIF value falls under the threshold. VIF based feature reduction has been used in literature before including [6], [13]. In this work, we focus on reducing the intra-modal collinearity for depression detection.

$$VIF_i = \frac{1}{1-R_i^2}$$

where R_i^2 is the explained variance of a regression model using the i^{th} feature as the dependent variable and the rest of the features as independent variables.

E. Shapley analysis

Shapley values[33] were introduced in game-theory to determine the contribution of individual players in cooperative games. More recently, it has been of interest to interest to the machine learning and explainable AI communities[8]. Given the i^{th} data instance with m features denoted by $X_i{}^m$, Shapley value for the j^{th} feature (equation-1) is defined as the weighted average of differences in predictions in the presence of the j^{th} feature and when it is marginalized. In practise, marginalization is achieved by using predictions from different subsets of features. Because of the intense (runtime increases exponentially with number of features) marginalization process, calculating Shapley value is computionally expensive. However, approximate Shapley values can be calculated using the SHAP (SHapley Additive exPlanations) framework[21]. These are referred to as the SHAP value.

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

where φ_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.

Local Interpretable Model-agnostic Explanations[32] (LIME) have been popular in explainable AI community to interpret model decisions. They approximate a given model using a linear model so that predictions from both models match (local accuracy) at least at the inputs. While LIME guarantees local accuracy, the SHAP framework improves to guarantee robustness to missing features (missing features have no impact on the contribution of the feature of interest) and consistency of features. This is achieved by kernelizing the traditional LIME model.

In this work, the SHAP values were used to rank the features in terms of their relative contribution to the task. Kernelized LIME was used to determine the SHAP values of features. We limit ourselves to Shapley analysis on multimodal features.

F. Classification setup

To enable interpretable features, following Alghowinem et al. [2] we use an SVM classifier to distinguish between mothers in the depressed and non-depressed groups in the feature space. All modalities except the LIWC were Z-score normalized, since they were designed to be implicitly normalized (each feature can take a value from 0 to 100%). Despite the recent success of deep learning models[19] and

TABLE I

Unimodal and Multimodal performance with no feature selection

	# features	ACC	PA	NA
Action Units	126	0.618	0.592	0.642
Face & Head Dynamics	143	0.594	0.384	0.702
Speech Behavior	29	0.534	0.553	0.510
Verbal	92	0.631	0.611	0.650
All modalities	390	0.671	0.662	0.679

their ability to learn task-relevant features, we refrain from using them over concerns of overfitting on the dataset. Additionally, we wish to inform developmental science by identifying theory-relevant features and an understanding of their relative contributions.

We explored SVM hyperparameters with linear and polynomial kernels and C value in the range of 10^{-5} to 10^{3} on a log-scale. The best hyperparameters for each experiment were determined with a gridsearch and five-fold cross-validation (CV) through gridsearchcv[29]. The best hyperparameters were used for Leave-One Subject-Out (LOO) cross-validation to report accuracy (ACC), positive agreement (PA) and negative agreement (NA) for all our experiments. Definitions of PA and NA can be found in Girard et al.[16].

IV. Results

In this section we present the results using unimodal and multimodal features as described in section-III. *All modalities* refers to early fusion of all unimodal features.

A. Depression Prediction

We evaluated how the unimodal features and multimodal all modalities perform at depression prediction without feature reduction. Features from different modalities described in section-III were used and the results are summarized as table-I. Among all unimodal features, the verbal features were the best performing modality. This was followed by the action units and face and head dynamics. Speech behavior modality contribution to the performance was only slightly better than a chance classifier. The early fusion of all modalities used outperformed the best unimodal features by 4% with an accuracy of 0.671.

B. Collinearity corrected features for depression prediction

a) VIF threshold: The choice of VIF threshold for collinearity correction was determined using a task-centric approach. We used a large range (0 to 100) for VIF threshold in order to be conservative with the number of features rejected in the first-step. We performed a LOO CV to determine the optimal VIF threshold for each configuration. For all modalities condition, we first performed collinearity correction for unimodal features and then used the resultant features with an early fusion strategy.

The thresholds for VIF determined using cross-validation (figure-1) reveal that the optimal collinearity among modalities differ. A low optimal threshold (VIF=20) for action unit features suggest that the corresponding 17 action unit features contain sufficient task-relevant information and further increases in collinearity led to a decrease in the performance.

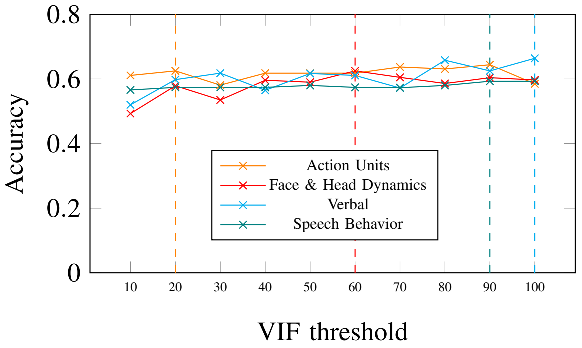

Fig. 1. Cross-validation accuracy of unimodal features at different levels of VIF threshold. The dashed vertical lines correspond to the best VIF threshold used for collinearity correction for each modality and experiments in table-II.

The face and head dynamics at VIF=10 (figure-1) had only 2 features, while they were less collinear, their efficacy at distinguishing depressed and non-depressed groups was poor. At optimal VIF=60 threshold, 29 features qualified suggesting that increasing collinearity could offer improvement in performance. On further increase at VIF=80, performance drops to 0.586. This trend shows that there is a collinearity-performance trade-off that could be explored. Verbal features had the highest VIF threshold. This could be because of the hierarchical nature of the features in the LIWC features (for example, the negative emotion feature overlaps with anxiety, anger and sadness but has some unique features). A higher VIF threshold allows for such collinear and task-relevant features to be used for prediction. The speech behavior features were robust to the changes in the VIF threshold though the optimal threshold was found to be 90.

Results from the last experiment presented the utility of different modalities in their original configuration of features. In this experiment we avoid features that were collinear. Table-II shows results after features from each modality with VIF exceeding the corresponding threshold were dropped. The unimodal performance improved across all modalities (between tables–I and II). This was particularly noticeable in speech behavior features where accuracy increased from 0.534 to 0.593. The performance of all modalities increased after accounting for collinearity, accuracy increased from 0.671 to 0.717.

In addition to performance improvements, feature reduction through collinearity correction were also observed. Comparing tables-I and II, we notice 1.5-7 times fewer features using VIF across modalities. Across all modalities, nearly a 4 times feature reduction was observed.

TABLE II

Collinearity corrected performance. All modalities is the early fusion of VIF reduced unimodal features

	VIF	# features	ACC	PA	NA
Action Units	20	17	0.644	0.620	0.667
Face & Head Dynamics	60	29	0.625	0.612	0.637
Speech Behavior	90	19	0.593	0.613	0.569
Verbal	100	40	0.664	0.658	0.671
All modalities	N/A	105	0.717	0.715	0.719

C. *Shapley analysis*

In Shapley analysis, we identified the top contributing features based on their SHAP values to further reduce the features used. We then performed classification using the top-k features and determined the subset of features that perform the best.

Avoiding collinearity demonstrated performance improvements and feature reduction in both unimodal and multimodal configurations, however, the necessity of VIF as an intermediate step against Shapley analysis with original features without collinearity correction was not established. To address this concern, we performed Shapley analysis on two configurations–the collinearity corrected early fusion of modalities (all modalities in table-II) and the early fusion of all modalities in their original configuration (all modalities in table-I) independently.

For each configuration, SHAP values were used to determine the feature contributions to the respective models. Features were then ranked based on their mean SHAP values–higher the value, more important the feature. To determine the best k, LOO CV was performed. Figure-3 shows the top-20 features for the collinearity corrected all modalities classifier. The performance trend against k for each configuration is shown as figure-2. As the number of features increase, the performance decreases after reaching the optimum, despite the feature relevance (i.e. ranked based on their relative contribution to the prediction). For Shapley on all modalities (no collinearity correction), much lower performance was observed initially. Adding more relevant features increased the performance with best performance at k=*85*. Both configurations achieved comparable best performance, the collinearity corrected Shapley approach achieved an accuracy of 0.777 with k=*15* features while the Shapley on original features achieved 0.782 accuracy with k=*85* features. We observe that collinearity correction results in a large reduction in features (nearly 6 times fewer features) over using Shapley analysis over collinear features.

TABLE III

Comparison of the best performing Shapley models

	All modalities	All modalities with collinearity correction
Features	85	15
ACC	0.782	0.777
PA	0.778	0.779
NA	0.787	0.773

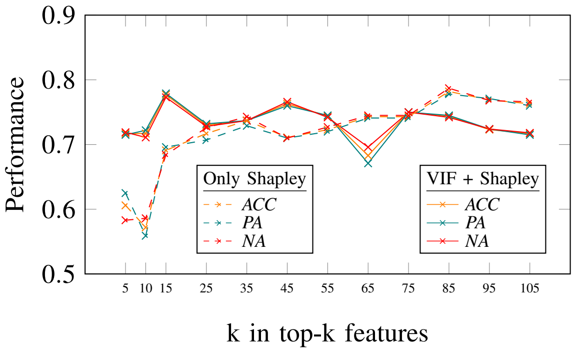

Fig. 2. Cross-validation performance of top-k features derived from Shapley analysis. The solid lines correspond to collinearity corrected all modalities and the dashed lines correspond to all modalities without collinearity correction.

V. Discussion

Among unimodal features the order (highest to lowest) of performance was verbal, action units, face and head dynamics and speech behavior. Face and head dynamics were better at identifying non-depressed class (high NA and low PA) over the depressed class. A similar trend was observed with action units and verbal features but the differences were smaller. Unlike other unimodal features, the speech behavior was better at depressed class prediction than the non-depressed class.

Early fusion of all modalities led to improvement in performance, and optimal collinearity threshold led to a large drop in the number of features across modalities. On an average about four times reduction in the number of features could be observed. The face and head dynamics based features have seen a five times reduction in the number of features. The speech behavior features were the least effected with a 1.5 times reduction and the largest improvement (6%) in performance across all metrics. The unimodal face and head dynamics evidenced a large improvement in PA (0.384 to 0.612) through collinearity correction and thus an increased sensitivity towards the depressed class. It also helped in reducing 143 features to 29 features. The biggest unimodal feature reduction was observed with action units where 7.9 times fewer features led to accuracy improvement from 0.618 to 0.644. Post feature reduction, using early fusion of modalities with 105 features led to 0.717 accuracy against its original configuration of 390 features with accuracy 0.671, nearly a four time reduction in the number of features. This demonstrates the utility of collinearity correction for prediction performance through

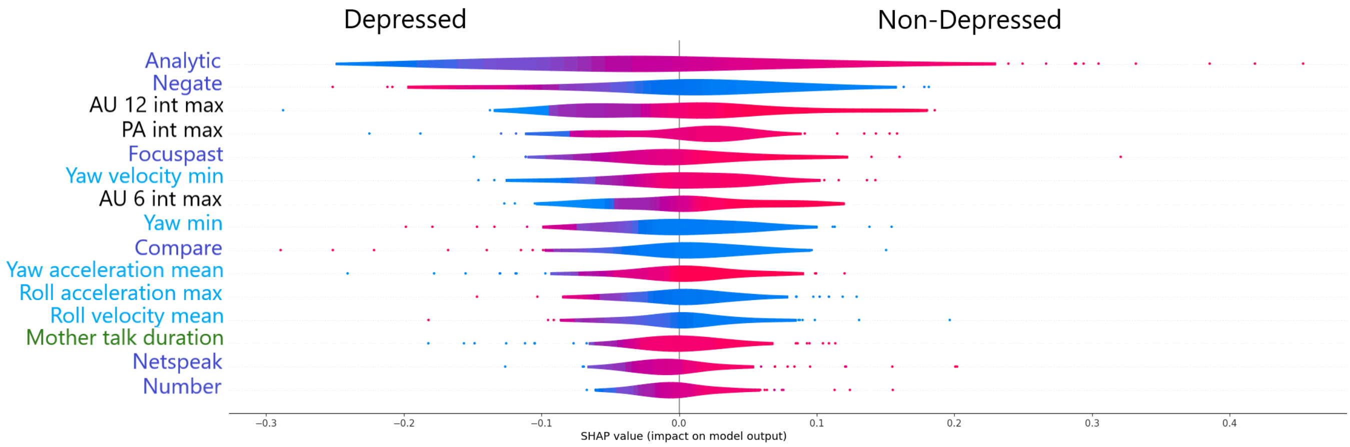

Fig. 3. Distribution of SHAP values for the top-15 collinearity corrected features as violin plots. The features along the y-axis are color coded by modality: LIWC, action units (AU), speech behavior (SB), and face and head dynamics (FHD). The x-axis denotes each feature's contribution, which may range from strongly negative (blue) to strongly positive (red).

TABLE IV

TOP-15 SHAPLEY RANKED FEATURES FROM ALL MODALITIES AFTER COLLINEARITY CORRECTION

Action Units	Face and Head Dynamics	Speech Behavior	Verbal
AU12 intensity	Yaw velocity	Mother talk duration	Analytic
PA (AU6+12) intensity	Yaw displacement		Negations (negate)
AU6 intensity	Yaw acceleration		Past events (focuspast)
	Roll acceleration		Comparisons (compare)
	Roll velocity		Informal language/humor (netspeak)
			Quantitative words (numbers)

feature reduction. Using the top-15 features from Shapley analysis on collinearity corrected features, we achieved 0.777 accuracy against 0.782 with 85 features through Shapley analysis on features without collinearity correction.

The top-15 features obtained through Shapley analysis on collinearity corrected early fusion of modalities can be found in table-IV. Non-verbal features (action units together with face and head dynamics) constituted more than 50% of the top-15 features. Given that the unimodal verbal features were the best performing modality (both with and without collinearity correction), equivalently they were the most (6 out 15 features) contributing modality towards the best performing multimodal model. Verbal features captured concepts such as analytic capabilities, focus on past events, negations, comparisons, humor and quantitative aspects of the spoken content. The analytic capabilities of depressed mothers were found to be lower than the non-depressed mothers. However, indicators such as usage of negations (no, not, never) and comparisons (better, best, after) were higher. Lower usage of informal language and humor (yup, haha, boo) in addition to referencing past events and agreements were found in the interactions between depressed mothers and their children. Focus on past was found to lower in depressed mothers over non-depressed mothers, similar observations were made in patient-therapist interactions in Dirkse et al.[10] where over the course of therapy for current depression a significant increased usage of references to past were encountered in depressed patients.

Base rate differences among action units were observed in previous studies for understanding current depression[14], [15], our analysis revealed differences in intensity level. The maximum intensity of AUs 12, positive affect, 6 and 10 (outside the 15 features) were found to lower in depressed mothers. This shows that the smiles (including the Duchenne smiles) among depressed mothers were of less intensity when compared to smiles in non-depressed mothers and intensity of action units should be investigated in addition to their occurrence. Only one of the top-15 features correspond to speech behavior. The total duration of mother‘s spoken activity was lower in depressed group than the non-depressed groups suggesting a reluctance among mothers to participant in a problem-solving task involving their children.

Some of our findings in families converge with previous findings in non-family contexts, such as interviews with a clinician or an avatar. In both the family interaction context and the interview context, head dynamics varied between depressed and non-depressed participants. Head dynamics in particular were similar in both contexts [2]. These included yaw displacement and roll dynamics (velocity and acceleration). This convergence is all the more striking given the differences between studies not only in context (family interaction vs. clinical interviews) but also in feature selection methods. We used VIF and Shapley; [2] used an aggregation of feature selection approaches. Another convergence was for speech features. In particular, speech of depressed participants was more likely to focus on past events [10], which is consistent with clinical observations of rumination in depression. With respect to facial expression, smile occurrence and intensity varied similarly in both our family interaction data and in clinical interviews. [15] found that smiles were less frequent or intense in depression. These findings suggest that unimodal and multimodal features of depression are robust to differences in context. Further work will be needed to test this hypothesis.

VI. CONCLUSION

Using a wide range of both verbal and nonverbal features, we achieved state of the art discrimination between mothers with and without depression. Our findings converge with ones previous in suggesting the importance of multimodal features and feature selection.

Reducing features in a principled way optimized performance. The findings extend previous ones in multiple respects. Depression detection was within families during a problem solving task (mothers and their adolescent offspring) rather than in clinical interviews; VIF and Shapley approaches to feature selection rather than alternative approaches were used; relative contribution of nonverbal and verbal modalities was revealed. Experiments explored relative contributions of unimodal and multimodal features. Through convergence with previous research, we found that multiple features are robust to varied differences in context and relationships.

Two limitations may be noted. One is the lack of prosody based features. Prosody has been shown to be an important modality offering depression related information as physical manifestation in speech production process and the speech outcome[1], [2], [7], [38]. Whether inclusion of prosody based features among the multimodal ones we considered would boost performance is an empirical question. It is possible that prosody might supplant some of the predictive power of other features rather than result in net boost. That is a question for further research. Another limitation is that only mother specific features were used for prediction. Given the evidence from psychology literature[24], [27] on differences in child behavior pertinent to parent depression, the directions for future work also includes using the child features both in conjunction with the mother features and separately to evaluate depression detection.

REFERENCES

[1] S. Alghowinem et al. Cross-cultural depression recognition from vocal biomarkers. In *Interspeech*, pages 1943–1947, 2016.

[2] S. M. Alghowinem et al. Interpretation of depression detection models via feature selection methods. *IEEE Transactions on Affective Computing*, 2020.

[3] W. R. Beardslee et al. Children of parents with major affective disorder: a review. *The American Journal of Psychiatry*, 1983.

[4] J. Cech and T. Soukupova. Real-time eye blink detection using facial landmarks. *Cent. Mach. Perception, Dep. Cybern. Fac. Electr. Eng. Czech Tech. Univ. Prague*, pages 1–8, 2016.

[5] J. F. Cohn et al. Multimodal assessment of depression from behavioral signals. *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2*, pages 375–417, 2018.

[6] T. A. Craney and J. G. Surles. Model-dependent variance inflation factor cutoff values. *Quality engineering*, 14(3):391–403, 2002.

[7] N. Cummins et al. A review of depression and suicide risk assessment using speech analysis. *Speech communication*, 71:10–49, 2015.

[8] A. Das and P. Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*, 2020.

[9] H. Dibeklioğlu et al. Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE journal of biomedical and health informatics*, 22(2):525–536, 2017.

[10] D. Dirkse et al. Linguistic analysis of communication in therapist-assisted internet-delivered cognitive behavior therapy for generalized anxiety disorder. *Cognitive behaviour therapy*, 44(1):21–32, 2015.

[11] I. O. Ertugrul et al. Crossing domains for au coding: Perspectives, approaches, and measures. *IEEE transactions on biometrics, behavior, and identity science*, 2(2):158–171, 2020.

[12] M. B. First and M. Gibbon. The structured clinical interview for dsm-iv axis i disorders (scid-i) and the structured clinical interview for dsm-iv axis ii disorders (scid-ii). 2004.

[13] G. S. Folli et al. Variable selection in support vector regression using angular search algorithm and variance inflation factor. *Journal of Chemometrics*, 34(12):e3282, 2020.

[14] J. M. Girard et al. Social risk and depression: Evidence from manual and automatic facial expression analysis. In *FG*, pages 1–8. IEEE, 2013.

[15] J. M. Girard et al. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing*, 32(10):641–647, 2014.

[16] J. M. Girard et al. Sayette group formation task (gft) spontaneous facial expression database. In *FG*, pages 581–588. IEEE, 2017.

[17] J. Gratch et al. The distress analysis interview corpus of human and computer interviews. Technical report, USC, 2014.

[18] A. Haque, M. Guo, A. S. Miner, and L. Fei-Fei. Measuring depression symptom severity from spoken language and 3d facial expressions. *arXiv preprint arXiv:1811.08592*, 2018.

[19] Y. LeCun et al. Deep learning. *nature*, 521(7553):436–444, 2015.

[20] M. C. Lovejoy et al. Maternal depression and parenting behavior: A meta-analytic review. *Clinical psychology review*, 20(5):561–592, 2000.

[21] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *NeurIPS*, 30, 2017.

[22] M. Morales et al. A linguistically-informed fusion approach for multimodal depression detection. In *Proceedings of the Fifth CLPsych*, pages 13–24, 2018.

[23] B. W. Nelson et al. 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):1513–1526, 2021.

[24] B. W. Nelson et al. Psychobiological markers of allostatic load in depressed and nondepressed mothers and their adolescent offspring. *Journal of Child Psychology and Psychiatry*, 62(2):199–211, 2021.

[25] I. Onal Ertugrul et al. Afar: A deep learning based tool for automated facial affect recognition. In *FG*. IEEE, 2019.

[26] W. H. Organization. Adolescent Mental Health. `https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health`, 2021.

[27] H. Orvaschel et al. Psychopathology in children of parents with recurrent depression. *Journal of abnormal child psychology*, 16(1):17–28, 1988.

[28] R. M. O'Brien. A caution regarding rules of thumb for variance inflation factors. *Quality & quantity*, 41(5):673–690, 2007.

[29] F. Pedregosa et al. Scikit-learn: Machine learning in Python. *JMLR*, 12:2825–2830, 2011.

[30] J. W. Pennebaker et al. The development and psychometric properties of liwc2015. Technical report, 2015.

[31] R. J. Prinz et al. Multivariate assessment of conflict in distressed and nondistressed mother-adolescent dyads. *Journal of applied behavior analysis*, 12(4):691–700, 1979.

[32] M. T. Ribeiro et al. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on KDD*, pages 1135–1144, 2016.

[33] L. S. Shapley. A value for n-person games. *Classics in game theory*, 69, 1997.

[34] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[35] M. Valstar et al. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *AVEC*, pages 3–10, 2013.

[36] M. Valstar et al. Avec 2014: 3d dimensional affect and depression recognition challenge. In *AVEC*, pages 3–10, 2014.

[37] L. Yang et al. Integrating deep and shallow models for multi-modal depression analysis—hybrid architectures. *IEEE Transactions on Affective Computing*, 12(01):239–253, 2021.

[38] Y. Yang et al. Detecting depression severity from vocal prosody. *IEEE transactions on affective computing*, 4(2):142–150, 2012.