Leveraging Shared and Divergent Facial Expression Behavior Between Genders in Deception Detection

Gazi Naven, Taylan Sen, Luke Gerstner, Kurtis Haut, Melissa Wen, Ehsan Hoque
Department of Computer Science
University of Rochester
Rochester, NY, United States
gnaven@ur.rochester.edu, [tsen, khaut, mehoque]@cs.rochester.edu, [lgerstn3,mwen2]@u.rochester.edu

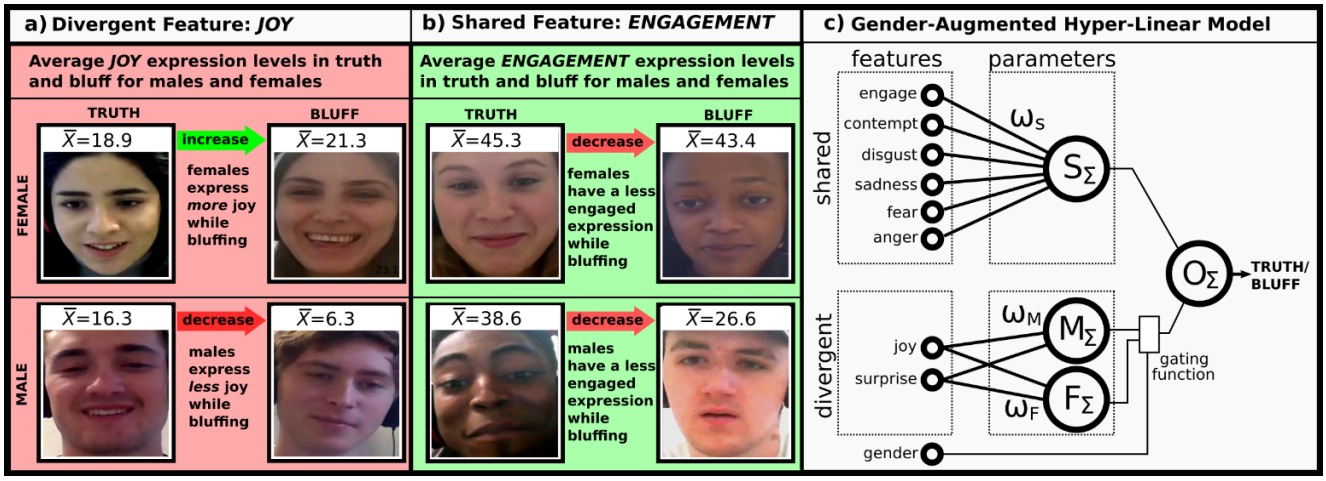

Fig. 1. Average Facial Expression Levels of Gender-Divergent and Gender-Shared Features During Truth and Bluffing in the GAHL Framework. Representative images showing average expression level on a 0-100 scale. a) Example divergent feature, *joy*. b) Example shared feature, *engagement*. c) The Gender-Augmented Hyper-Linear Facial Expression Model has both *shared* parameters ω_S (for which it is more beneficial to pool male and female training data) as well as separate male and female *divergent* parameters ω_M , ω_F (which are necessary to support non-linear complexity across genders).

***Abstract*— While facial expression behavior has been understood as mostly universal between the genders, recent research has highlighted important differences, including the expression of smiles and surprise. Despite such gender differences, studies involving facial expression often have limited sample sizes such that splitting the data set in half to train separate male and female models has been untenable. In order to leverage gender divergent complexity in facial expression models while also using a full dataset to train shared behaviors, we developed GAHL: the Gender-Augmented Hyper-Linear model. GAHL selectively increases non-linear model complexity with regards to gender divergent features. Using both simulated data and data from a study of facial expressions during deception (N=80, >6 hours), we demonstrate that when the facial expression data set size is in the range of N<75, GAHL outperforms several mainstream machine learning models including logistic regression, decision tree, and SVM with polynomial and radial basis function kernels.**

I. Introduction

A common trait of many dyadic behavioral studies is having small data size (i.e. N < 500). However, in order to exploit complex machine learning models (i.e. deep learning), large datasets (N > 1,000 and much larger) are typically required. As the bias-variance tradeoff famously dictates, using models which are too complex for a given data set leads to higher generalization error [5], even with the application of various regularization approaches [17]. This problem has been shown to be exacerbated by high dimensionality in the number of features [6]. Linear models have often been used with small data sets, and linear models with L1 regularization (i.e. LASSO regression) have demonstrated ability to handle large dimensions [12,13]. However, purely linear models are often just too simple to model the complexity of many problems.

One such source of complexity lies in the interpretation of facial expression differences between men and women [1,2]. Prior research has shown that men and women tend to express certain facial expressions associated with emotion at different levels depending on the situation due to emotion regulation. In an analysis of expressions of participants watching videos with affective content, McDuff et al. found that women tend to express more inner brow raise (AU1) and smile (AU12) than men [2] (where "AU" refers to the Facial Action Coding System "action unit" number [22,23]). Dimberg, et al., in

This research was supported in part by grant W911NF-15-1- 0542 and W911NF-19-1-0029 with the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO), and the National Science Foundation NRT-DESE #1449828.

investigating differences in male and female responses to viewing pictures with angry and happy faces, found that women showed more pronounced facial reactions, especially to happy faces [7]. While these findings may support a stereotype that women are more expressive than men [1], several studies have identified that for some facial expression types, women show less expression than men [2]. For example, McDuff et al., found that women express less brow furrow (AU4) [2], a facial action unit which has been associated with anger [3]. Similarly, Evers et al. found evidence that not only are women less likely to show expressions of anger, but that they are also more likely to suppress such expressions [3]. Hence, these studies show that there are inherent differences in the way certain facial action patterns are expressed between genders.

The nexus of low data set sample size with facial expression analysis is particularly true of studies involving deception, where N is often < 200 [14,16,19]. While studies have investigated deceptive communication using a vast array of diverse features and methods [27,28,29], facial expression analysis has been one of the most prominent areas of focus over the years. Ekman et al. considered facial expressions as a form of unintentional "emotional leakage" that provides clues of one's deception [12]. Ekman also showed that expressions of joy are associated with deception under a psychological theory called duping delight [19]. Duping delight is the premise that deceivers take delight in lying to another person, especially when there is an audience to the deceptive behavior [19]. Engagement has been found to be important in past literature for deception as well. Dunbar et al showed that synchrony also described as engagement was found to be a major factor for honest participants in a dyadic study [21]. However, there has been no previous research demonstrating that engagement is expressed at different levels between males and females. A multimodal approach was taken by Rosas et al. who considered both verbal and non-verbal behavior, including manually labeled facial expressions and hand movements. Their analysis identified that the most influential facial features of deception were frowning (AU4, AU7, and AU17 which is closely associated with anger) and brow raiser (AU1, AU2 which are closely associated with surprise) [22]. We look to further this past research of facial expression analysis to detect deception by leveraging both the differences and similarities of automated facial expression analysis between males and females.

In this paper, we present the GAHL (Gender-Augmented Hyper-Linear) facial expression model framework shown in Fig. 1, for data sets with low data sample size and high individual variance. GAHL selectively adds nonlinearity for specific traditional emotion-based facial features which have gender divergence. The GAHL model maintains two sets of parameters, those that are shared for both genders, and those that are separately maintained for males and females. The GAHL model uses the full data set to train shared parameters but allows nonlinear gender-specific complexity to exist regarding facial expression features which have been demonstrated to have gender-specific behavior. We apply the GAHL model to predict when participants in a small communication game study (N=80) are being honest or deceptive. Our contributions are summarized as follows:

- Development of the Gender-Augmented Hyper-Linear facial expression model which adds limited nonlinear complexity for gender-divergent facial expressions.
- Characterization of the superiority of the GAHL model over several standard machine learning models for a range of low data set sizes (N < ~100) using both simulated data as well as a dyadic interrogation data set involving honest and bluffing witnesses (N=80).
- Identification that the facial expressions associated with joy, surprise, and engagement are relevant to detecting deception, with joy and surprise demonstrating gender divergence, and engagement being shared between genders.

II. Methods

In this section we first describe the GAHL framework and model implementation. We then describe how the simulation data was generated followed by an overview of how the deception data was collected with its associated deception game protocol. The facial feature extraction methods are described, followed by the bootstrapping resampling methodology used to evaluate performance with various input training set sizes for GAHL and other classification methods used throughout this study. While much of the description is written in the context of the deception detection classification task, the GAHL model is intended to be broadly applicable to a wide variety of behavioral classification tasks.

A. The GAHL Framework

1) Feature Type Determination (divergent/shared)

Shown in Fig. 1 is a high-level description of the GAHL model framework. The first step of the GAHL framework involves identifying which of the facial expression input features are expected to be gender-divergent (Fig. 1a) vs. gender-shared (Fig. 1b) for the desired classification task (e.g. deception detection). This distinction can be made for each feature using a combination of domain knowledge and statistical analysis of the data. Domain knowledge most preferably would indicate gender divergence for the specific classification task at hand (i.e. which facial expression features have different meanings, or different meaning levels, for males and females regarding the classification task). Compared to specific domain knowledge, which may be rare, general domain knowledge, which provides non-task specific evidence of facial expression divergence between genders, may also be useful in determining divergent features. Statistical testing may be used either together with or in lieu of domain knowledge to individually test whether a given feature is gender-divergent or gender-shared for a given task. The Mann Whitney U test [9] may be used to test whether a given feature has different median values between the different

classification groups (i.e. truthful vs. deceptive speakers). We will use each of these feature type determination methods in categorizing divergent and shared features for the task of deception detection.

2) The Gender Augmented Hyper-Linear Model

Shown in Fig. 1c is a high-level graphical description of the mathematical data flow in the GAHL model, which is specifically described in equations 1-2 below

$$O(X) = \sum_{\forall i \in S} \omega_i^S x_i + \sum_{\forall i \in D} [\omega_i^F x_i * g_i + \omega_i^M x_i * (1 - g_i)] + \omega_o \quad [1]$$

$$O(X) = \begin{cases} TRUTH : x \geq 0 \\ BLUFF : x < 0 \end{cases} \quad [2]$$

where O represents the GAHL model's final summing node output, S represents the set of shared features, D represents the class of divergent features, ω^S represents the shared weights, ω^M represents the male weights, ω^F represents the female weights, g, a gender variable, which is 1 for female data points and 0 for males, and finally, ω_0 represents the summing offset (logistic regression intercept)

The GAHL model's input consists of the divergent features (X_i, i ∈ D) and shared input features (X_i, i ∈ S) separated into two groups of inputs as shown. The product of the shared features X_i, i ∈ S and the shared weights ω^S are summed in the circular node S_Σ. The divergent features are multiplied by two sets of weights, first the male weights, ω^M, which are then summed in the M_Σ node. Similarly, the divergent features are also multiplied by a set of female weights ω^F and then summed in the F_Σ node. The outputs of the M_Σ and F_Σ nodes both go to a gating function, which selectively allows only one output to pass as controlled by the gender input. More specifically, if the gender is female, then only the F_Σ node output is allowed to pass onto the final summing stage. Alternatively, if the gender for a given data point is male, the M_Σ node output is allowed to pass to the final summing node. The S_Σ node output always passes to the summing node regardless of gender.

It should be noted that the gating function does not need to be controlled by binary gender. Rather the model supports the use of non-binary (i.e. a gender value between 0 and 1). The resulting decision surface of the GAHL model is shown in Fig. 2. Effectively, the gating function introduces the helical nonlinear surface that cannot be achieved by a plain linear model alone (even if gender is added as an input). The complex nonlinearities associated with the gender variable are only introduced to the features which have been greedily identified as divergent. It should be noted that the complexity of this model (in terms of degrees of freedom) is linear in the number of features. More specifically, the total number of model parameters (degrees of freedom) is two times the number of divergent features and one times the number of shared features (plus one for the offset)

The parameters of the GAHL model were trained with stochastic gradient ascent. It is worth pointing out that the shared weights, ω^S, will be affected by all training data, whereas the male weights ω^M are only updated by male data points, and the female weights ω^F are updated by only female data points

B. Simulation Data

Simulated data used to evaluate GAHL is generated as a collection of data points with gender labels (male/female) and outcome labels (truth/bluff) over a two-dimensional input feature space. The two input features are simply referred to as the "shared feature" and the "divergent feature". Each of the features were designed to have a log-linear effect on the probability of the outcome being truth/bluff. The shared feature provides the same influence on the truth/bluff determination for both male and female datapoints. Alternatively, the gender-divergent feature is designed to have opposite effect on the outcome variable depending upon the gender. More specifically, the shared feature has a direct positive effect on the likelihood that both male and female data points are truth (i.e. the larger the shared-feature, the more likely that a data point is truth, regardless of gender). In contrast, as the divergent feature gets larger, male data points are more likely to be truth, but female datapoints are less likely to be truth. These rules were used to generate a data set (N = 2000) with the distributions shown in Fig. 3. Note how for both males and females, the frequency of truth (yellow) increases as the shared feature increases. Alternatively, as the divergent feature increases, the male truth frequency increases while the female truth frequency decreases. It is important to note that the ideal separating classification line for the male and female data are perpendicular. Even if gender is added as a feature, a linear model alone would never be able to implement two perpendicular classification boundaries. This simulated data is used to evaluate the performance of GAHL as well as several other models.

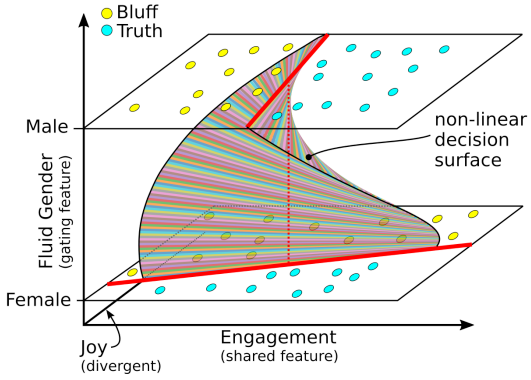

Fig. 2. Representation of GAHL Model Decision Surface.

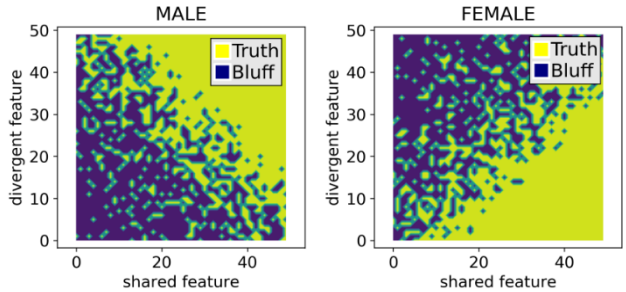

Fig. 3. Simulated Data Distributions of Truth and Bluff Among Males and Females Over the Space of a Shared and Divergent Feature.

C. Deception Game Protocol

Participants were recruited to play a deception game with another participant over video chat. The deception game is a customized version of the Automatic Dyadic Data Recorder framework [9]. Participants were recruited from a university community and were randomly assigned the role of either witness or interrogator. The game began with the witness being instructed to memorize an image shown by the web application for a 60 second viewing period. The web application then randomly instructed the witness to either bluff or tell the truth regarding the image to the interrogator. Then the interrogator was directed by the web application to ask the witness several predetermined questions about the image to determine if the witness was bluffing or telling the truth. Following these questions there was a short period of time where the interrogator was encouraged to ask the witness their own questions.

Each participant played up to eight rounds, four as an interrogator and four as a witness. The witnesses were motivated to get the interrogator to believe them by providing a monetary bonus each round. If the witness bluffed and the interrogator believed them, they received $20, and if they told the truth and the interrogator believed them, they received $10. The interrogators were motivated to correctly determine whether the witness was lying or not by providing them with a monetary bonus each round as well. They received $10 if they were able to correctly determine if the witness was bluffing or telling the truth. In addition, participants were notified ahead of time that one of the eight rounds they played would be randomly selected to be a "high stakes round", in which the bonus amount increased to $50. The total number of unique dyads in the study was N=80 in which 36 male witnesses (14 truth, 22 bluff) and 44 female witnesses (20 truth, 24 bluff).

D. Automated Facial Expression Extraction

Facial expressions were extracted from the relevant questioning segments of the raw video study data using the Affdex tool [10]. Affdex uses an SVM based model trained over one million facial expression videos containing affective-rich content. We use expression level values for joy, fear, sadness, surprise, disgust, contempt, and engagement from this tool [22, 23].

Using this tool, we were able to extract automated facial features for over 6 hours of video with their respective emotion-associated expression levels labeled. The emotion-associated expression features are measured on a scale of 0 to 100. The use of emotion-associated features vs. individual facial action units was motivated by the conjecture that deception and honesty are more directly related to emotional states rather than the contraction of specific facial muscle groups [31].

To validate the Affdex tool we used 327 images of participants expressing a given emotion from the Cohn-Kanade CK+ Database [30]. The CK+ database provides categorical emotion labels, i.e. each image is given a single emotion label out of the set of: Happy/Joy, Anger, Disgust, Fear, Sadness, Surprise, and Contempt. Affdex provides multidimensional output in that each emotion category has its own independent numerical output. In order to convert Affdex numerical output into categorical data, the CK+ emotion category with the maximum output is selected as the label. The maximum-category Affdex emotion label matched with the hand labeled CK+ emotion label 74.0 percent of the time. This task was also done by a human labeler with an accuracy of 75.2 percent, which is comparable to the Affdex accuracy.

E. Classification

The classification accuracy for the GAHL model was compared to several models including: logistic regression with gender as an input and both L1 and L2 regularization, a M/F Log. Reg. (two separate logistic regression models, one for female data and one for male data), Naive Bayes, and SVM with radial basis function and polynomial (degree 2) kernels. For each of the classifiers and each of the training set sizes, the models were rerun using 1000 random data splits to reduce randomness in the results. During these runs ½ of the training data was used as a development set to determine the ideal hyper parameters. The models were tested across different training set sizes to further demonstrate the effectiveness of GAHL and show the "sweet spot" of training set sizes where GAHL is most useful.

III. Results

In this section we discuss the results from the simulated dataset and from the deception dataset. We present the analysis of the statistical differences within each dataset and show how the GAHL model was able to improve the accuracy for the classification tasks within each dataset. Furthermore, for the linear models the weights of the features are displayed.

A. Simulation Data

Shown in Fig. 4 are the bootstrap resampled classification results of the tested models on the simulation data. In the range of training set sizes tested, the GAHL model demonstrates the best performance. The gender-separated logistic regression model ("M/F Log. Reg.") approaches the same performance level as the GAHL model at a training set size of ~105 samples (74.3% vs.74.1%). The maximum performance improvement between GAHL and the next best performing M/F Log. Reg. model occurs at a training set size of 18 with the respective accuracies of 68% and 61%. As demonstrated in Fig. 4, the logistic regression model performs markedly worse than the GAHL model at all tested training set sizes. For reference, the performance of a featureless model is trained, with the sole parameter of the predicted truth to bluff ratio (i.e. the "Prob" model)

B. Deception Data

In this section, we outline results from the deception data. First, we describe the divergent/shared feature type assignment based on both domain knowledge and statistical test results. Second, we show how GAHL and other mainstream models perform on this dataset. Finally, we show weights of each feature from the linear models to show how the divergent features in GAHL play a major role.

1) Domain Knowledge for Feature Type Selection

As introduced in the background section of the introduction, here is evidence indicating that smiling (joy) and engagement are expected to be relevant in deception detection for both males and females [18,19,23]. Additionally, anger and surprise have been associated with deception without gender divergence [22]. Further, independent of deception, research has suggested that there are divergent facial expression behaviors in males and females regarding smiling and anger [2, 24]. This evidence suggests that the joy and anger features should be divergent.

2) Statistical Findings

Table I shows that in the deception dataset, female bluffers and truth-tellers differed in expression levels of joy, and surprise with uncorrected p-values from the Mann Whitney U test of 0.040, and 0.048 respectively. Specifically, female bluffers show higher levels of joy with average expression levels of 23.1 compared to 18.8 of female truth-tellers. Expressions of surprise by females were found to be expressed in higher levels by truth-tellers with average levels of 10.7. In males, engagement was found to be expressed in higher expression levels with an average of 38.6 (p=0.031). It should be noted that after Bonferroni multiple test correction, none of these p-value are statistically significant [25].

3) Classifier Performance

We compare and show our performance findings from the GAHL model using joy and surprise as the divergent features, a M/F Log. Reg. model, and linear and non-linear models that use gender as a feature. Fig. 5 shows the performance results for the models explained above in the Methods section, on the deception dataset.

Across the deception results, the GAHL classification model performs best once the training set size reaches 25. The model's best performance is with accuracy of 61.5% on the test set using a training set size of 66. This is the highest accuracy among all the models across all the training set sizes. At this same training size, the M/F Log. Reg. achieves accuracy of 57.2% and the next closest competitor is the SVM model using a polynomial function kernel with gender as an additional feature, which achieves an accuracy of 54.1%. The largest difference between GAHL and the M/F Log Reg. model is also when the training size is 66, with a difference of 4.3%. Fig. 5 shows the effectiveness of GAHL, when there is low data a sweet spot exists where GAHL can perform better than all the other models including the M/F Log. Reg. model. In this instance the sweet spot occurs once the training set size is greater than or equal to 25. From Fig. 5, it can be seen clearly that the GAHL model with its divergent features outperforms the linear and non-linear models with features of emotions and gender included as another feature.

TABLE I: DIFFERENCES BETWEEN TRUTHFUL AND DECEPTIVE WITNESSES GROUPED BY GENDER.

Gender	Feature	Buff	Truth	Cohen's d	p-value
Female	**joy**	**23.1**	**18.8**	**-0.396**	**0.040**
	surprise	**6.5**	**10.7**	**0.667**	**0.048**
	engagement	43.4	45.3	**0.130**	0.282
Male	joy	6.3	16.3	0.801	0.070
	surprise	6.7	5.1	-0.294	0.391
	engagement	**26.6**	**38.6**	0.836	**0.031**

4) Statistics from Linear and GAHL Model features

While statistical analysis using the Mann-Whitney U test in Table I allow us to view features in isolation, p-values of features from the linear models' weights provides deeper insights into the relative importance of features. In this regard we consider significance test for each feature in the three model configurations: i) a logistic regression model with gender as an additional feature, ii) M/F Log. Reg. models for males and females, iii) the GAHL model with its associated divergent and shared features [20].

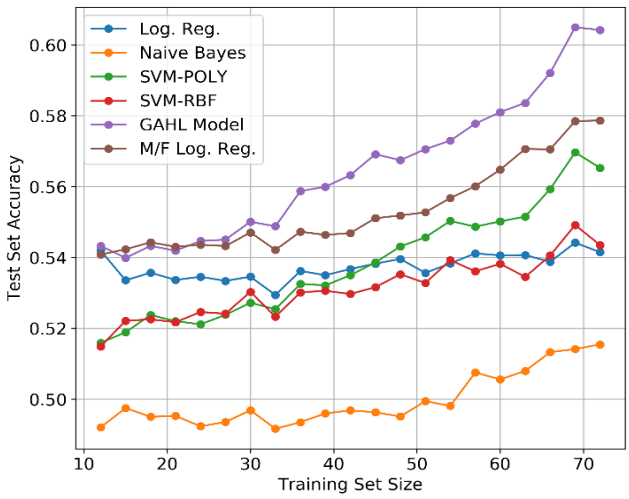

Fig. 4. Comparative Performance of GAHL Model vs. Training Set Size

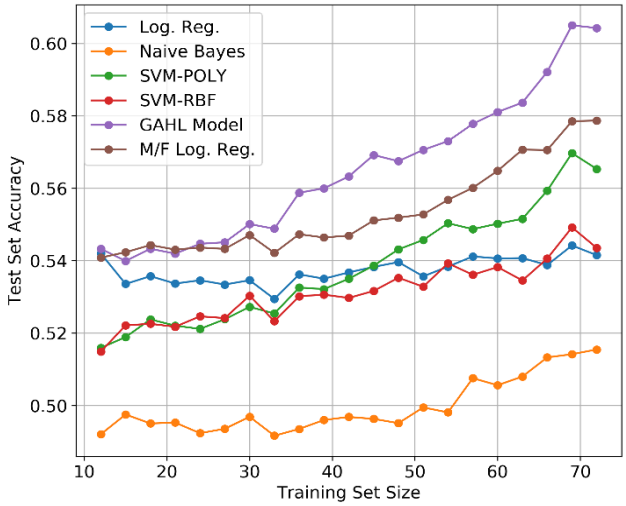

Fig. 5. Comparative Performance of GAHL Model with Varying Deception Data Set Size.

TABLE II: SIGNIFICANCE OF FEATURES FROM LINEAR MODELS

Model	Feature	p-value	weight
Logistic Regression	joy	0.312	-0.090
	surprise	0.553	-0.049
	engagement	0.207	0.075
M/F Log. Reg.: Male	joy	0.305	-0.394
	surprise	**0.016**	-1.522
	engagement	**0.019**	0.864
M/F Log. Reg.: Female	joy	0.650	-0.059
	surprise	0.803	0.085
	engagement	0.746	0.030
GAHL	joy_m	0.312	-0.116
	joy_f	**0.030**	-0.239
	surprise_m	**0.029**	-0.304
	surprise_f	0.461	-0.069
	engagement	**0.024**	0.168

In Table II, notice that joy, surprise, and engagement have been highlighted among other Affdex features used. These features appeared to be significant in at least one of the models and so we show it to draw comparisons. We find that none of the features are significant when a simple logistic regression has been used. In the case of Gender Separated Model where the logistic regression was trained on males and females separately, surprise and engagement (p-value of 0.016 and 0.019 respectively) are significant. For the case of GAHL, it is interesting to find that joy_f (joy weights for females), surprise_m (surprise for males), and engagement were also significant (p-values of 0.03, 0.029, and 0.024 respectively)

IV. DISCUSSION

The results provide several insights regarding both gender divergence and deception detection. To a large degree, the divergences suggested by the domain knowledge references discussed in the introduction are validated not only by the performance of the GAHL model, but also in the statistical significance of the model weights. The performance curves over varying training set sizes in Fig 4 and 5 also highlights the importance of models tuned to divergent feature complexity in low data set sizes.

A. Divergent Features

As mentioned in the Methods section divergent feature selection is carried out with a combination of domain knowledge and statistical test of difference in the dataset for the given features. In the Results section, Table I shows us that males and females have differences in expressions of joy, engagement, and surprise when they are telling the truth or lying. We find evidence that support divergence of joy and surprise, but not for engagement. The duping delight theory, which suggests increased feelings and expressions of joy in deceivers [19,8], is supported in the findings involving female witness participants, but not the male participants. The reason why the duping delight theory was not shown on average in the male participants may lie in the fundamentally different expression levels of smiles exhibited among males and females [2].

Interestingly, as shown in Table I, only the truthful female witnesses express more surprise as identified by [2]. We surmise that the majority of data collected by [2] did not involve deception, as individuals were voluntarily responding to an online video and did not have incentive to be deceptive. This is most likely the main reason the data among bluffing witnesses does not agree with the findings identified by [2].

In addition to surprise, inner brow raiser has also been shown to be associated with fear [2]. However, our statistical results did not show any significant divergence in our deception dataset between males and females with regards to fear. Similarly, anger, sadness, disgust, and contempt also showed no statistically significant divergence between males and females. Our finding that facial expressions of engagement is not divergent among males and females is consistent with past literature which identified no differences between males and females regarding the facial action units associated with engagement (i.e. outer brow raiser (AU2), and lip corner depressor (AU15) [2].

It is important to note that our findings may be limited to the particular cultural demographic of our study participants. While Ekman et al. and others have demonstrated pan-cultural elements of facial expressions of emotions, studies, such as Jack, et al., have also shown exceptions based on culture, notably Western and Eastern cultures [29]. Indeed, Ekman et al.'s findings even recognized cultural differences in masking emotions [28]. However, it should be emphasized that the main contribution of this paper is not the particular findings with this dataset, but rather the introduction of a hyper-linear model that can leverage both divergent and shared features between two classes in small datasets. Thus, it is possible that some facial expressions are universal across cultures, while others are divergent. This invites the application of a modified GAHL model in which features are shared and divergent not across genders, but across some other group, such as cultural groups (such as the Westerner and Easterner groups identified by Jack et al.) [29]. Thus, in this paper, we recognize that in any given dataset there may be facial expressions which are shared (i.e. universal) as well as facial expressions which are divergent (i.e. exceptions to universality). We demonstrate how our proposed GAHL model, can maximize performance by providing limited nonlinear complexity regarding non-universal differences when they exist.

B. Classifier Statistical Tests in Deception Data

We present significance tests to provide further empirical evidence about existence of feature divergence and how GAHL was able to leverage that information in its predictive ability. This analysis was done on i) a logistic regression model with gender as an additional feature, ii) M/F Log. Reg. models for males and females, iii) the GAHL model with its associated divergent and shared features because they are linear in nature and are the top three models in terms of accuracy performance on the deception data.

In Table II, notice that there were no significant features from the logistic regression model. It warrants our argument about this model being too simple to perform well in the complexity that exists in facial expression interpretation in the deception dataset. Introducing the M/F Log. Reg. model gives us some complexity. Specifically, in the male model, we find

that surprise and engagement are significant features in predicting honest and deceptive behavior. Surprise was chosen as one of the divergent features for GAHL. While engagement showed up to be significant for divergence, we did not find enough domain knowledge to confirm its divergence. Hence, it is not surprising that the M/F Log Reg. model is relying on these features during prediction. Finally, for GAHL, we find that both joy and surprise are significant predictors. It is significant particularly for joy in females, and for surprise in males. Hence, we find both divergent features contributed to increasing the complexity of the model. It is also interesting to note that engagement which was not a significant feature for the simple logistic regression became significant when used as a shared feature in GAHL. Engagement has a coefficient weight of 0.17 in GAHL which shows that it is associated with honest behavior. In fact, engagement's significance has been found to be important in past literature too for deception where synchrony or engagement was found to be a major factor for honest participants in a dyadic study [21].

C. Classifier Performance

When there is a low data count, the GAHL model is able to outperform models that simply include gender as a feature and a separate model for each gender. In the model that includes gender only as an additional feature, insufficient importance is put on the differences between males and females. The two separate models for males and females are not able to perform as well as the GAHL model when there is a low data count. Splitting the data into two distinct groups based on gender lowers the training data count even further making it harder for the individual models to be trained. Using the GAHL model, it allows us to have separate weights for features that are determined to be gender divergent. The model also allows us to combine features that are shared, or non-divergent features. This maximizes the amount of data we can use to train a model that will generalize well past the training set.

We see the difference the data count makes on which model will work best in the deception data and the simulated data. The GAHL model achieves higher performance compared to all the other models for the deception data (N = 80), as shown in Fig. 4. This adds evidence to the idea that our model will be able to perform at its best when there is a low data count but will not substantially outperform training two different models based on gender when there is sufficient data available. Furthermore, the ability for two separate models to be able to perform nearly as well as the GAHL model when there is sufficient data adds to the argument that there are major differences in the expression of emotions between males and females.

D. Non-Binary Gender Data

In a cross-cultural study, Fontanella et al. found the presence of a multi-dimensional gender identity [27]. They find that their claims are similar to the general consensus of the research community on gender identity theory that there does exist fuzzy boundaries between male and female gender identities. Our aim is to improve our dataset by incorporating a self-identified gender scale that can range between maleness and femaleness, that will replace the binary gender data. This could be easily incorporated in the GAHL model by using this gender scale instead of the binary gender feature as shown in Fig. 2. This means that the divergent feature weights will be given weighted importance instead of absolute importance that is proportionate to the gender scale.

V. Conclusion

The GAHL model was successfully applied to deception detection, demonstrating a clear region of data set size in which the performance surpasses that of all other models. Perhaps more importantly, evidence of the existence of gender divergent facial feature meanings is also introduced. We are hopeful that other researchers may use the GAHL model and may further benefit from a nuanced approach to investigate context-specific facial expression meaning in additional domains. We also hope that a model like GAHL can be applied to other domains with different gating features such as age, instead of gender, where that feature dictates certain patterns of differences across other features.

Acknowledgment

This research was supported in part by grant W911NF-15-1- 0542 and W911NF-19-1-0029 with the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO), and the National Science Foundation NRT-DESE #1449828.

References

[1] Fabes, Richard A., and Carol Lynn Martin. "Gender and age stereotypes of emotionality." Personality and social psychology bulletin 17, no. 5 (1991): 532-540.

[2] McDuff, Daniel, Evan Kodra, Rana el Kaliouby, and Marianne LaFrance. "A large-scale analysis of sex differences in facial expressions." PloS one 12, no. 4 (2017): e0173942.

[3] Fischer, Agneta H., and Catharine Evers. "Anger in the context of gender." *International handbook of anger*. Springer, New York, NY, 2010. 349-360.

[4] McLean, Carmen P., and Emily R. Anderson. "Brave men and timid women? A review of the gender differences in fear and anxiety." Clinical psychology review 29, no. 6 (2009): 496-505.

[5] Abu-Mostafa, Yaser S., Malik Magdon-Ismail, and Hsuan-Tien Lin. Learning from data. Vol. 4. New York, NY, USA:: AMLBook, 2012.

[6] Bishop, Christopher M. Pattern Recognition and Machine Learning. S.l.: Springer-Verlage New York, 2016.

[7] Dimberg, Ulf, and Maria Petterson. "Facial reactions to happy and angry facial expressions: Evidence for right hemisphere dominance." *Psychophysiology* 37.5 (2000): 693-696.

[8] Sen, Taylan, Md Kamrul Hasan, Minh Tran, Matthew Levin, Yiming Yang, and Mohammed Ehsan Hoque. "Say cheese: Common human emotional expression set encoder and its application to analyze deceptive communication." In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 357-364. IEEE, 2018.

[9] T. Sen, M. K. Hasan, Z. Teicher, and M. E. Hoque, "Automated dyadic data recorder (ADDR) framework and analysis of facial cues in deceptive communication," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 1, no. 4, p. 163, 2018.

[10] Mann, Henry B., and Donald R. Whitney. "On a test of whether one of two random variables is stochastically larger than the other." *The annals of mathematical statistics* (1947): 50-60.

[11] McDuff, Daniel, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana el Kaliouby. "AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit." In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, pp. 3723-3726. ACM, 2016.

[12] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE, 2016, pp. 1–10.

[13] Ekman, Paul, and Wallace V. Friesen. Facial action coding system: Investigator's guide. Consulting Psychologists Press, 1978.

[14] Ekman, Paul, and Wallace V. Friesen. "Nonverbal leakage and clues to deception." Psychiatry 32, no. 1 (1969): 88-106.

[15] Dreber, Anna, and Magnus Johannesson. "Gender differences in deception." Economics Letters 99.1 (2008): 197-199.

[16] Cody, Michael J., and H. Dan O'Hair. "Nonverbal communication and deception: Differences in deception cues due to gender and communicator dominance." Communications Monographs 50.3 (1983): 175-192.

[17] Ho, Shuyuan Mary, and Jonathan M. Hollister. "Guess who? An empirical study of gender deception and detection in computer-mediated communication." *Proceedings of the American Society for Information Science and Technology*50.1 (2013): 1-4.

[18] Niyogi, Partha, and Federico Girosi. "On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions." Neural Computation 8, no. 4 (1996): 819-842.

[19] P. Ekman, Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition). WW Norton & Company, 2009.

[20] L. Ten Brinke and S. Porter, "Cry me a river: Identifying the behavioral consequences of extremely high-stakes interpersonal deception." Law and Human Behavior, vol. 36, no. 6, p. 469, 2012

[21] Peng, Chao-Ying Joanne, Kuk Lida Lee, and Gary M. Ingersoll. "An introduction to logistic regression analysis and reporting." The journal of educational research 96.1 (2002): 3-14.

[22] Dunbar, N. E., Jensen, M. L., Tower, D. C., & Burgoon, J. K. (2014). Synchronization of Nonverbal Behaviors in Detecting Mediated and Non-mediated Deception. Journal of Nonverbal Behavior, 38(3), 355-376

[23] Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., & Burzo, M. (2015, November). Deception detection using real-life trial data. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (pp. 59-66). ACM.

[24] Yu, X., Zhang, S., Yan, Z., Yang, F., Huang, J., Dunbar, N. E., ... & Metaxas, D. N. (2015). Is interactional dissynchrony a clue to deception? Insights from automated analysis of nonverbal visual cues. IEEE transactions on cybernetics, 45(3), 492-506.

[25] McDuff, D. (2017, October). Smiling from adolescence to old age: A large observational study. In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 98-104). IEEE.

[26] Cabin, Robert J., and Randall J. Mitchell. "To Bonferroni or not to Bonferroni: when and how are the questions." *Bulletin of the Ecological Society of America* 81, no. 3 (2000): 246-248.

[27] Fontanella, L., Maretti, M., & Sarra, A. (2014). Gender fluidity across the world: a Multilevel Item Response Theory approach. Quality & Quantity, 48(5), 2553-2568.

[28] Ekman, Paul, and Dacher Keltner. "Universal facial expressions of emotion." Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture (1997): 27-46.

[29] Jack, Rachael E., et al. "Facial expressions of emotion are not culturally universal." Proceedings of the National Academy of Sciences 109.19 (2012): 7241-7244.

[30] Lucey, Patrick, et al. "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression." 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. IEEE, 2010.

[31] Stel, Mariëlle, and Eric van Dijk. "When do we see that others misrepresent how they feel? detecting deception from emotional faces with direct and indirect measures." Social influence 13.3 (2018): 137-149