Dataset Bias in Deception Detection

Ara Mambreyan
Department of Engineering
University of Cambridge
United Kingdom
Email: ara.mambreyan99@gmail.com

Elena Punskaya
Department of Engineering
University of Cambridge
United Kingdom
Email: op205@cam.ac.uk

Hatice Gunes
Dep. of Computer Science & Technology
University of Cambridge
United Kingdom
Email: hatice.gunes@cl.cam.ac.uk

Abstract*—With the advances in Machine Learning, lie detection technology gained significant attention. In recent years, several multi-modal techniques achieved as high as 99% accuracy results using the Real-life Trial dataset with only 121 data points. This led to considerable media hype and research interest in lie detection with machine learning. In this paper, we analyze the effect of dataset bias in deception detection. More specifically, we train a classifier to predict the sex of the identity appearing in the video. On a test data point, we use the sex predictor to predict sex which we use as a proxy for predicting deception, predicting *lie* for females and *truth* for males. This lie predictor simulates a classifier that uses nothing but dataset bias. Nevertheless, we find that the performance of this biased classifier is comparable to those of state-of-the-art papers. More specifically, when using IDT features, our biased classifier achieves 64.6% and 59.3% AUC while a classifier trained normally on truth/lie labels achieves 57.4% accuracy and 69.3% AUC. We perform similar experiments on the Bag-of-Lies dataset and show that it too is biased with respect to sex. In addition, we apply the state-of-the-art techniques on an unbiased dataset and show that their performance is no better than chance. Our experiments strongly suggest that the results of recent deception detection techniques can be explained by the bias inherent in the datasets.

I. Introduction

A flawless lie detection technology has the potential to revolutionize the justice system, combat terrorism and minimize the spread of fake news. Many methods and devices were researched over the decades, including the polygraph for measuring physiological responses and fMRI for brain-scanning. However, both the polygraph and the fMRI have been shown to have high error rates [1]–[3]. Recently, new machine learning methods for deception detection were researched which use various modalities such as text, audio and video [4]–[14]. Some of these techniques achieved extremely good results in deception detection. In particular, metrics as high as 92.21% AUC [8], 96.13%, [9], 97.0% [10] and 99% [11] accuracies were reported using only 121 data points [7].

The success of these papers resulted in significant attention in both popular media and academic settings alike. Outlets such as The Guardian [15], Financial Times [16] and WIRED Magazine [17] commented on the rise of AI Lie Detectors, the latter describing them as a "Black Mirror World". In an interview with the media outlet Futurism in 2018, one of the authors of [8] even claimed "we could be just three to four years away from an AI that detects deception flawlessly by reading the emotions behind human expressions" [18].

While a flawless lie detector would be an important invention, we are not aware of any research papers which quantitatively analyze potential issues with current techniques such as dataset bias and fairness. Deception Detection is a sub-field of Affective Computing where dataset bias and unfairness have been studied considerably [19]–[21]. It has been shown that affective computing tasks are often biased with respect to sensitive attributes such as sex[1], race and age. Furthermore, results can be highly overestimated if the distribution of labels from these attributes is imbalanced [19]–[21] which not only wastes valuable research time but also advertises false expectations of these technologies to the media and governments.

Given the ethical implications of AI Lie Detectors, we set out to investigate potential issues that could have caused the almost perfect results of modern deception detection techniques. In particular, we perform experiments to analyze dataset bias in the Real-life Trial dataset [7] used in the above-mentioned techniques and the Bag-of-Lies dataset [14]. Finally, we apply the state-of-the-art techniques to the Bag-of-Lies dataset and the Miami University Deception Detection dataset [22]. The latter has no dataset bias. The main contributions of this paper are as follows:

(1) We show that two deception detection datasets, Real-life Trial dataset [7] and Bag-of-Lies dataset [14], are biased, particularly for the sex attribute, and should not be used in isolation.

(2) We quantify the impact of the sex bias for these datasets by training a classifier to predict sex and using it as a proxy for predicting deception.

(3) We show that the techniques achieving almost perfect results on the Real-life Trial dataset [7] are no better than chance when applied on the Bag-of-Lies dataset [14] and the Miami University Deception Detection dataset [22].

The rest of this paper is structured as follows. Section II provides background literature on deception detection and dataset bias. Section III describes the datasets. Section IV describes the experiments and Section V presents the results. Section VI provides a high-level discussion on the findings and presents further quantitative and qualitative arguments of the inadequacy of current state-of-the-art papers. Finally, section VII concludes the paper.

[1] Throughout this paper, all occurrences of the terms *male* and *female* refer to biological sex.

II. Background Literature

A. Deception Detection

Datasets

Different types of datasets were used for deception detection including simulated lab experiments [14], [22], [23], games [24]–[27] and real-life [7]. Datasets from simulated lab experiments are usually acquired by asking the subjects to tell facts about themselves [22], describe an image they are shown [14], participate in mock crime scenes [23], etc. In recent years, attention in deception detection shifted toward more realistic scenarios. To our knowledge, the Real-life Trial dataset [7] is currently the only publicly available dataset from real-life scenarios. The dataset was used in recent state-of-the-art papers which achieved extremely high metrics [8]–[13].

The datasets include modalities such as audio, video [7], [22], manually annotated facial expressions [7], [26], gaze [14]. Often, a multi-modal approach is used [14].

Techniques

Both hand-crafted feature extraction and automatic feature extraction with deep learning were attempted for video classification of deception detection [8]–[14] with varying degrees of success. In this paper, we mainly focus on the papers which achieved extremely high results. In particular, we discuss and analyze two of the most cited papers: [8] which used Improved Dense Trajectory features and [9] which used a 3D-CNN neutral network for the video features. However, most of our analysis also applies to other papers which achieved similarly high results.

1) Improved Dense Trajectories: IDT (Improved Dense Trajectories) are hand-crafted features that were shown to achieve state-of-the-art results on various action recognition tasks [28]. Since deception detection is, in some sense, an action recognition task of humans, IDT features were extracted and classified using various machine learning algorithms in [8], achieving a 77.31% AUC score. When combined with other modalities, 92.21% AUC score was achieved.

2) 3D-CNN: 3D-CNN is an extension of 2D-CNN where convolution is achieved by convolving a 3D kernel [29]. 3D-CNN has been shown to achieve state-of-the-art results in human action recognition tasks [29]. Hence, it was attempted for deception detection classification from videos, achieving 93.08% accuracy when using video features and 96.14% when combining with other modalities [9].

Besides IDT and 3D-CNN, other techniques were used for deception detection in recent works. For instance, in [10], a derivative of the popular two-stream network [30], along with meta-learning and adversarial learning, was used achieving a 93.16% accuracy when using only video features and 97.00% when combining it with other modalities.

The above-mentioned techniques were only applied to the Real-life Trial dataset. For the Bag-of-Lies dataset, LBP (Local Binary Patterns) are computed for the video modality achieving 55.26% accuracy. For the gaze modality, 57.11% accuracy is achieved using Random Forest and 62.71% when fusing all modalities.

B. Dataset Bias

Dataset bias is a well-known problem in Machine Learning and, especially, Computer Vision. Algorithms trained on biased datasets have high unfairness scores [31], [32] and can highly overestimate the actual metrics of the intended task by learning incidental properties of the dataset [33]. For example, if, hypothetically, all subjects who lied in a given dataset wore black shirts and all subjects who were truthful wore white shirts, a simple classifier could easily achieve 100% accuracy by detecting the colour of the shirt of the subject; yet, clearly, this classifier would have nothing to do with actual deception detection.

A more concrete example of dataset bias occurred in Criminality-from-Face classifiers where the aim is to predict whether the person appearing in the picture is a criminal. Almost perfect scores were achieved in this task; however, it was identified that the reason for the impressive results was not because the algorithms actually learned how to distinguish criminality from face but because the data points of positive and negative labels were from different domains (such as social media and actual police mug shots) and the algorithms simply learned to distinguish the domains [34].

Another example occurred in the Kinship Detection research where significantly high results were achieved; however, it was shown that some of the datasets of kinship detection were biased as images of related subjects (such as father and son) were cropped from the same original image [35] which caused the algorithms to learn this incidental property. It is clear from the examples of Kinship Detection and Criminality-from-Face research that impressive results can be achieved solely by the algorithms learning to exploit the biases of the dataset. Hence, the research question we address in this paper is the following: are almost perfect metrics of the state-of-the-art papers in deception detection the result of dataset bias and flaws in experimental designs or do AI Lie Detectors actually work?

III. Datasets

We use three datasets in our experiments: Real-life Trial dataset [7], Bag-of-Lies dataset [14] and Miamy University Deception Detection dataset [22].

A. Real-life Trial dataset

All recent papers achieving almost perfect results used the Real-life Trial dataset. The dataset consists of 121 videos from real-life courtroom hearings with 61 lies and 60 truths from 56 identities. The videos were shot in an unconstrained setting with significant variations in pose, illumination and size. Figure 1 shows several screenshots taken from the dataset.

The features for each data point consist of a video (including audio), transcript and manually annotated micro-expressions. The latter is a 39-length feature vector where each entry of the vector is binary-valued and specifies whether a particular micro-expression appeared on the subject's face throughout the video. The dataset is highly imbalanced from identities as several identities dominate the dataset.

Figure 1. Screenshots from the Real-life Trial dataset videos [7]. Notice the large variations in the settings.

The lie percentage from sexes is shown in Table I. The large difference in lie percentages is likely coincidental as the four identities with the highest number of data points in the dataset, totalling 50 data points, follow the sex bias trend significantly affecting the bias of the whole dataset.

Table I
REAL-LIFE TRIAL DATA DISTRIBUTION FROM SEX.

Sex	# of Points	Lie %
Female	76	64.5%
Male	45	26.7%

Figure 2. Screenshots from Bag-of-Lies dataset [14]. Top row: truths; bottom row: lies.

B. Bag-of-Lies dataset

The dataset consists of 325 recordings with 162 lies and 163 truths from 35 identities. The recordings were shot in a lab setting where the subjects were asked to describe an image shown to them on a screen. Multiple modalities are provided including video (with audio), Gaze and Electroencephalogram (EEG). The gaze data includes features such as fixation points and pupil sizes. Figure 2 shows several screenshots taken from the dataset. The lie percentage from sexes is shown in Table II.

While not as large as in the Real-life Trial dataset, there is still considerable sex bias present in the Bag-of-Lies dataset, likely coincidental.

Table II
BAG-OF-LIES DATA DISTRIBUTION FROM SEX.

Sex	# of Points	Lie %
Female	94	61.7%
Male	231	45.0%

The dataset defect for both the Real-life Trial dataset and the Bag-of-Lies dataset means that if the features of the dataset are correlated with sex an algorithm may achieve statistically significant results without actually learning any patterns for lying. In Section IV, we assess this effect on the results by simulating an algorithm that uses nothing but the sex bias to predict lying.

C. Miami University Deception Detection dataset

The dataset consists of 80 identities; 20 Black Females, 20 Black Males, 20 White Females and 20 White Males. Each identity has 4 data points - 2 truths and 2 lies - and, hence, there are 320 data points in total. The videos were shot in a lab setting where subjects were asked to lie or tell a truth about their social relationships. Figure 3 shows several screenshots taken from the dataset.

Figure 3. Screenshots from Miami University Deception Detection dataset [22]. Top row: truths; bottom row: lies.

All videos of the same subject were shot in the exact same setting[2]. This is critical for our experiments as there are no sensitive attributes the algorithms could exploit to achieve high results.

IV. EXPERIMENTS

All experiments were written in **Python**; the deep learning architecture for the 3D-CNN was implemented using **PyTorch** [36] and classical Machine Learning classifiers were written using **scikit-learn** [37]. The code for Sections IV-A and IV-B1 is published in [38].

A. Experiments with Real-life Trial dataset

To compare our results with those of other papers, we run 10-fold cross-validation similarly to [8]–[10]. To avoid the algorithm from degenerating to person re-identification, all data points of the same identity reside in the same fold.

[2] One subject wears a jacket in only two of her data points but these include 1 truth and 1 lie and, hence, do not contribute to the dataset bias.

Thus, a subject who is in the training set does not appear in the validation set. In addition, we noticed that there is a large variance in the metrics depending on how the folds were sampled. Hence, we ran cross-validation for a number of iterations, sampling new folds randomly for each iteration subject to the above-mentioned constraints; the exact number of cross-validation iterations is mentioned in each experiment.

Finally, to simulate a sex-biased algorithm, we train a machine learning classifier to predict sex. The exact classifiers used are mentioned in the next two sub-sections. On a test data point, the classifier predicts sex and uses this as a proxy for predicting deception – predicting *lie* if the predicted sex is *female* and vice-versa. Figure 4 shows this process.

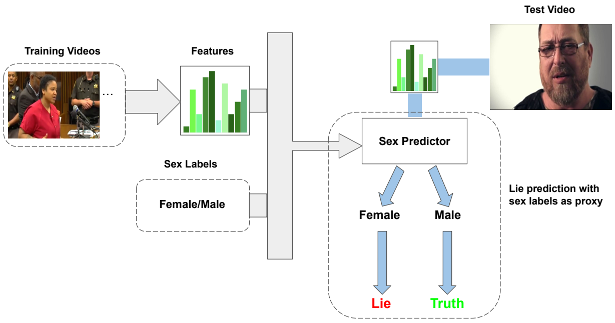

Figure 4. Illustration of the experiment that simulates a sex-biased algorithm to predict deception.

To have a baseline for the simulated classifier, we also train another classifier in the normal way with truth/lie labels. This is the same classifier as those of state-of-the-art papers but we also run this to minimize implementation differences with the simulated classifier.

1) Manually Annotated Micro-expression features: Firstly, we use micro-expressions to assess sex bias in the Real-life Trial dataset. Cross-validation is run for 25 iterations for the classifiers Linear SVM, Logistic Regression, K-Nearest Neighbours and 5 iterations for Kernel SVM, Random Forest, Adaboost and MLP (Multilayer Perceptron). The mean value of cross-validations is reported for each classifier.

A hyperparameter search was performed for all classifiers. For Kernel SVM, we used RBF functions and tuned the regularization parameter and the variance of the kernel functions. For Random Forest, the maximum depth was tuned. For Adaboost, the number of estimators and the learning rate were tuned and we used a one-depth Decision Tree as the weak learner. For MLP, we used one hidden layer and tuned the number of hidden cells along with the L2 regularization parameter.

2) IDT features: As described in Section II, [8] used IDT features with Fisher Vector encodings to predict deception. In this experiment, we directly use the extracted features and encodings provided by [8] to minimize the differences between the implementations. [8] pruned 15 selected videos claiming "the pruned videos have either significant scene change or human editing". To replicate their experiment design, we prune the same set of videos. In this experiment, we only use a Linear SVM classifier which achieved the highest score in [8] when using only the video modality. We tune the regularization hyperparameter **C**.

[8] did not report the large variance of the results from the sampling of the folds. In contrast, we run cross-validation for 25 iterations, randomly sampling new folds for each run of cross-validation. To show the large range of results, we report minimum, mean (average) and maximum metrics for all 25 iterations of cross-validation. In addition, we report both Accuracy and AUC metrics. Finally, similarly to Section IV-A1, we run two experiments to assess sex bias.

B. Experiments with Bag-of-Lies dataset

1) Gaze features: We follow the exact same procedure as in Section IV-A except with 3-fold cross-validation to replicate the experimental design of [14]. Specifically, all data points of the same identity reside in the same fold. Similarly to Section IV-A, we run two experiments for assessing sex bias in the Bag-of-Lies dataset; in the first experiment, we train the classifier on sex labels as illustrated in Figure 4 and, in the second experiment, we train in the normal way with truth/lie labels.

As features, we use the gaze data provided by the dataset and, similarly to [14], extract fixation points and pupil size. The duration of the fixation, x and y coordinates are constructed for the top 20 fixations ranked by their duration. In addition, we use the number of fixations, average pupil size, standard deviation of the pupil size and pad 1 to each feature vector to get a 64-dimensional feature vector.

Random Forest and MLP classifiers are used and the same set of hyperparameters are tuned as in Section IV-A1. The average of 5 iterations of cross-validation is reported.

2) IDT features: To assess whether the IDT technique works outside of the Real-life Trial dataset, we attempt the technique on the Bag-of-Lies dataset following the exact same procedure as in Subsection IV-A2. However, we only train on truth/lie labels and don't use sex labels in this experiment.

C. Experiments with Miami University Deception Detection dataset

As the Miami University Deception Detection dataset is an unbiased dataset, we attempt two of the state-of-the-art techniques, IDT and 3D-CNN, on this dataset.

We run three separate experiments, training on truth/lie labels in the normal way. We use a train-test split and run 10-fold cross-validation on the train split to tune the hyperparameters. Similarly to the Real-life Trial dataset, we ensure that the same identity does not appear in both training and testing.

IDT features: We use the same set of classifiers and tune the same hyperparameters as for manually annotated micro-expressions in Section IV-A1.

Cross-domain: We also train on the Real-life Trial dataset, using IDT features, and test on the Miami University Deception Detection.

Table III
THE ACCURACY SCORES ACHIEVED WITH DIFFERENT CLASSIFIERS FOR THE REAL-LIFE-TRIAL DATASET USING MICRO-EXPRESSIONS. L: TRAINED WITH LIE/TRUTH LABELS. S: TRAINED WITH SEX LABELS.

ACC	LR	KNN	L-SVM	K-SVM	RF	MLP	AB
L	76	74	75	74	76	74	**76**
S	60	**64**	62	61	61	63	59

3D-CNN: We attempt various architectural adjustments and hyperparameter settings with a 3D-CNN network.

V. Results

A. Experiments with Real-life Trial dataset

1) Manually Annotated Micro-expression features: Table III shows the results of the experiments with the Real-life Trial dataset when using manually annotated micro-expressions. The accuracies achieved when training the classifiers with truth/lie labels roughly agree with those of other papers [7], [8].

We can see that the classifiers trained on sex labels achieve statistically significant results. Hence, a classifier that uses no information about deception whatsoever achieves seemingly impressive results in deception detection. Furthermore, it highlights that the results achieved by classifiers trained on the Real-life Trial dataset, even when only using manually annotated micro-expressions, are not reliable. This has not been reported in any of the papers using this dataset [7]–[13].

2) IDT features: Table IV shows the results of the experiments with the Real-life Trial dataset using IDT features. Firstly, we note that the maximum AUC score obtained from 25 runs of cross-validation when trained on truth/lie labels is 74.2% compared to the 77.31% AUC reported in [8]. Since we used the feature encodings provided by [8], the difference between implementations was minimal. We believe the different metrics achieved can be attributed to a different set of data points used; in their paper, [8] claims that 15 selected videos were pruned and provide a list of pruned videos that we used for our experiments as mentioned previously. However, the publicly available code for [8] shows that only 5 videos were pruned.

Secondly, the minimum, mean and maximum of 25 cross-validations are shown in Table IV. There is a significant difference in the metrics depending on the sampled folds. For instance, there is an 8.9% AUC difference between the minimum and maximum obtained when trained on truth/lie labels. These random errors from folds were not reported in the state-of-the-art papers.

Finally, the classifier trained on sex labels achieves comparable results to the one trained on truth/lie labels when using IDT features. For the accuracy metric, the classifier trained on sex labels achieves better scores than a classifier trained on truth/lie labels – a difference of 7.2% between the means. For the AUC, the classifier trained on sex labels is worse than the classifier trained on truth/lie labels by 10.0%.

This result clearly shows that using nothing but the bias of the dataset it is possible to achieve worse AUC scores and better accuracy scores than a model trained with truth/lie labels. This is significant and strongly suggests that the model which achieved extremely high metrics in [8] is exploiting the dataset bias.

Table IV
THE MINIMUM, MEAN AND MAXIMUM ACCURACY AND AUC SCORES ACHIEVED FOR THE REAL-LIFE TRIAL DATASET USING LINEAR SVM WITH IDT FEATURES. L: TRAINED WITH LIE/TRUTH LABELS. S: TRAINED WITH SEX LABELS.

ACC	Min	Mean	Max
L	52.4	57.4	61.0
S	59.1	64.6	68.8

AUC	Min	Mean	Max
L	65.3	69.3	74.2
S	55.0	59.3	65.8

Table V
ACCURACY ACHIEVED FOR THE BAG-OF-LIES DATASET. L: TRAINED WITH LIE/TRUTH LABELS. S: TRAINED WITH SEX LABELS.

ACC	RF	MLP
L	59.3	55.5
[14]	57.1	53.5
S	53.5	54.4

B. Experiments with Bag-of-Lies dataset

1) Gaze features: Table V shows the accuracy results obtained with the gaze features along with the accuracies obtained in [14]. We first note that the accuracy results obtained training on truth/lie labels are higher than those of [14]. This difference is likely due to minor differences in feature engineering, implementation and random errors.

Secondly, the classifier trained on sex labels achieves statistically significant results. This shows that even a classifier that uses gaze features can exploit the underlying dataset bias of the Bag-of-Lies dataset. We chose the gaze features to demonstrate the dataset bias for their simplicity. However, we note that other features, especially video and audio, are likely to produce higher results with the sex-biased classifier as those features are likely more correlated with sex.

2) IDT features: Using IDT features similarly to the Real-life Trial dataset on the Bag-of-Lies dataset produces results no better than chance.

C. Experiments with Miami University Deception Detection dataset

As in the case of the Bag-of-Lies dataset, none of the experiments we performed on the Miami University Deception Dataset achieved better than chance results. This provides evidence that state-of-the-art techniques are unable to predict deception. However, we note the limitations of this experiment; namely, the Bag-of-Lies dataset and the Miami University Deception Detection dataset are shot in a lab setting where the stimuli are presumably lower than in the Real-life Trial dataset.

VI. Discussion

Lie detection is a highly complex task for which humans' performance is only slightly above chance [39]. At the same time, state-of-the-art machine learning-based lie detection methods achieve almost perfect accuracies [8]–[13] using only 121 data points. The accuracies achieved by these methods exceed the state-of-the-art classification accuracies for the popular ImageNet dataset [40] which includes 14+ million images even though it is a much simpler task for humans than lie detection [40]. At first glance, the metrics achieved by the state-of-the-art lie detection methods are surprising and may even seem revolutionary. However, although the potential applications of lie detection are vast and important, we believe the recent techniques claiming almost perfect results can be explained by dataset bias.

In Sections V-A1 and V-A2, we showed that the Real-life Trial dataset used for state-of-the-art papers has significant sex bias and it is possible to achieve high scores by exploiting the dataset bias. Although we only analyzed sex bias, there might be other types of dataset biases such as race, age, background in the video, quality of the video, clothing, pose, glasses, hair colour, etc. that might be correlated with deception in the Real-life Trial dataset. If a classifier learned to exploit these biases it could easily achieve high metrics while not learning any actual patterns for deception. However, the state-of-the-art papers did not consider these limitations and conduct sensibility checks to test dataset bias. We further showed that the Bag-of-Lies dataset has considerable underlying bias which the algorithms could exploit to achieve high results. Hence, in deception detection research, dataset bias is not unique to the Real-life Trial dataset.

Furthermore, experiments in Section IV-B2 and IV-C indicate that these techniques are no better than chance when applied to the Bag-of-Lies dataset and the Miami University Deception Detection dataset [22]. At the same time, state-of-the-art papers have not attempted their techniques on any dataset other than the small and imbalanced Real-life Trial dataset [7]. Hence, as in Kinship Detection research and Criminality-from-Face classifiers, described in Section II-B, there is no good reason to believe that the state-of-the-art techniques in lie detection actually learn to distinguish deception and not just dataset bias.

Besides the issues of dataset bias, there were other issues not emphasized in the previous sections. Importantly, no test split was used in the experiments of the state-of-the-art papers. This means the algorithm could overfit on the validation set and not generalize [41]. This is even more relevant since the validation splits used were very small and there was no mention of limiting the number of models and hyperparameters used. In addition, in Section V-A2, we showed that there is a significant range of the metrics, both accuracy and AUC, from the data distribution among the folds by performing 25 iterations of cross-validation and reported minimum, mean, and maximum values. These random errors were not reported in any of the state-of-the-art papers.

VII. Conclusion

Due to the serious ethical implications of AI Lie Detectors and their hype in media and industry alike, we investigated the dataset bias of the Real-life Trial dataset and the Bag-of-Lies dataset. For both datasets, we confirmed that they include significant sex bias as females in the datasets lie more. Furthermore, we demonstrated that machine learning algorithms trained on these datasets, some of which reported metrics of 92.21% AUC [8], 96.14% [9], 97.00% [10] and 99% [11] accuracy, are unreliable as they can exploit the incidental properties of the datasets and appear as if they learned to distinguish deception. Finally, we tried some of the state-of-the-art techniques, which achieved almost perfect results on the Real-life Trial dataset, on the Bag-of-Lies dataset and the Miami University Deception Detection dataset and showed that they achieve no better than chance results.

Although our main attention in this paper was to quantify the impact of dataset bias, we note that these models are discriminatory and biased with respect to sensitive attributes such as sex, race and age. In the best-case scenario, techniques trained on biased deception detection datasets could highly overestimate the capabilities of lie detection and, in the worst case, be used as tools for discrimination.

In the future, we strongly recommend that the Real-life Trial dataset and the Bag-of-Lies dataset are not used in isolation. In addition, any new datasets created should attempt to have minimal dataset bias, for instance, by ensuring that all subjects have the same percentage of truths and lies and all videos of the same subject are shot in the same setting similar to the Miami University Deception Detection dataset [22], described in Section III-C. For future research, we also recommend that researchers in the field be wary of potential dataset biases or issues that could cause very high metrics. Sensibility checks need to be performed before concluding that a technique learns to distinguish deception.

We hope our findings will save time for researchers who work with the assumption that current state-of-the-art papers in lie detection are valid and achieve almost perfect accuracies. Finally, we hope that our findings will recalibrate expectations of AI Lie Detection technology for both industry and governments.

Data Access Statement

This study involved secondary analyses of pre-existing datasets. All datasets are described in the text and cited accordingly. Licensing restrictions prevent sharing of the datasets.

Acknowledgements

Hatice Gunes is supported in part by the European Union's Horizon 2020 Research and Innovation program project WorkingAge under grant agreement No. 826232, and in part by the EPSRC under grant ref. $EP/R030782/1$. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any Accepted Manuscript version arising.

References

[1] A. Vrij, *Detecting Lies and Deceit: The Psychology of Lying and the Implications for Professional Practice.* Wiley, 2001.

[2] T. Gannon, A. Beech, and T. Ward, *The Use of the Polygraph in Assessing, Treating and Supervising Sex Offenders*. Wiley, 2009, ch. Risk Assessment and the Polygraph.

[3] M. J. Farah, J. B. Hutchinson, E. A. Phelps, and A. D. Wagner, "Functional mri-based lie detection: scientific and societal challenges." *Nature reviews Neuroscience.*, vol. 15, no. 2, pp. 123–131, 2014.

[4] R. Mihalcea and S. Pulman, "Linguistic ethnography: Identifying dominant word classes in text." *CICLing, Springer.*, p. 594–602, 2009.

[5] J. Pennebaker, M. Francis, and R. Booth, "Linguistic inquiry and word count (LIWC): LIWC2001," vol. 71, 2001.

[6] J. Hirschberg, S. Benus, J. Brenier, F. Enos, S. Friedman, S. Gilman, C. Gir, G. Graciarena, A. Kathol, and L. Michaelis, "Distinguishing deceptive from non-deceptive speech," in *Proceedings of Interspeech 2005 - Eurospeech*, 2005, p. 1833–1836.

[7] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, p. 59–66.

[8] Z. Wu, B. Singh, L. S. Davis, and V. S. Subrahmanian, "Deception detection in videos," in *AAAI*, 2018.

[9] G. Krishnamurthy, N. Majumder, S. Poria, and E. Cambria, "A deep learning approach for multimodal deception detection." *CICLing, Springer*, 2018.

[10] M. Ding, A. Zhao, Z. Lu, T. Xiang, and J. Wen, "Face-focused cross-stream network for deception detection in videos," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7794–7803, 2019.

[11] N. Carissimi, C. Beyan, and V. Murino, "A multi-view learning approach to deception detection," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 2018, pp. 599–606.

[12] M. Gogate, A. Adeel, and A. Hussain, "Deep learning driven multimodal fusion for automated deception detection," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1–6.

[13] S. Venkatesh, R. Ramachandra, and P. Bours, "Robust algorithm for multimodal deception detection," in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019, pp. 534–537.

[14] V. Gupta, M. Agarwal, M. Arora, T. Chakraborty, R. Singh, and M. Vatsa, "Bag-of-lies: A multimodal dataset for deception detection," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*, 2019, pp. 83–90.

[15] A. Katwala, "The race to create a perfect lie detector – and the dangers of succeeding," https://www.theguardian.com/technology/2019/sep/05/the-race-to-create-a-perfect-lie-detector-and-the-dangers-of-succeeding, 2019, accessed: 20-01-2022.

[16] C. Hodgson, "AI lie detector developed for airport security," https://www.ft.com/content/c9997e24-b211-11e9-bec9-fdcab53d6959, 2019, accessed: 20-01-2022.

[17] M. Harris, "The lie generator: Inside the black mirror world of polygraph job screenings," https://www.wired.com/story/inside-polygraph-job-screening-black-mirror/, 2018, accessedL 20-01-2022.

[18] D. Galeon, "A new AI that detects "deception" may bring an end to lying as we know it," https://futurism.com/new-ai-detects-deception-bring-end-lying-know-it, 2018, accessed: 20-01-2022.

[19] T. Xu, J. White, S. Kalkan, and H. Gunes, "Investigating bias and fairness in facial expression recognition," in *ECCV Workshops*, 2020.

[20] S. Yan, D. Huang, and M. Soleymani, "Mitigating biases in multimodal personality assessment," in *Proceedings of the 2020 International Conference on Multimodal Interaction*. New York, NY, USA: Association for Computing Machinery, 2020, p. 361–369.

[21] S. Li and W. Deng, "A deeper look at facial expression dataset bias," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[22] E. Lloyd, J. Deska, and K. Hugenberg, "Miami university deception detection database." *Behav Res.*, vol. 51, p. 429–439, 2019.

[23] J. Gonzalez-Billandon, A. M. Aroyo, A. Tonelli, D. Pasquali, A. Sciutti, M. Gori, G. Sandini, and F. Rea, "Can a robot catch you lying? a machine learning system to detect lies during interactions," *Frontiers in Robotics and AI*, vol. 6, p. 64, 2019.

[24] S. Demyanov, J. Bailey, K. Ramamohanarao, and C. Leckie, "Detection of deception in the mafia party game," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, p. 335–342.

[25] F. Soldner, P. Verónica, and R. Mihalcea, "Box of lies: Multimodal deception detection in dialogues," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

[26] J. E. Loy, H. Rohde, and M. Corley, "Cues to lying may be deceptive: Speaker and listener behaviour in an interactive game of deception." *Journal of Cognition*, vol. 1, no. 1, p. 42, 2018.

[27] G. Chittaranjan and H. Hung, "Are you awerewolf? detecting deceptive roles and outcomes in a conversational role-playing game," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 5334–5337.

[28] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A robust and efficient video representation for action recognition." *International Journal of Computer Vision.*, vol. 119, no. 3, p. 219–238, 2016.

[29] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[30] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc., 2014.

[31] M. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, 2017, p. 4069–4079.

[32] J. Buolamwini and T. Gebru, "sex shades: Intersectional accuracy disparities in commercial sex classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, vol. 81. New York, NY, USA: PMLR, 23–24 Feb 2018, pp. 77–91.

[33] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1521–1528.

[34] K. W. Bowyer, M. C. King, W. J. Scheirer, and K. Vangara, "The "criminality from face" illusion," *IEEE Transactions on Technology and Society*, vol. 1, no. 4, pp. 175–183, 2020.

[35] M. B. López, E. Boutellaa, and A. Hadid, "Comments on the "kinship face in the wild" data sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 11, pp. 2342–2344, 2016.

[36] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[37] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[38] A. Mambreyan, "lie-detector-code," https://github.com/AraMambreyan/lie-detector-code, 2022, accessed: 21-01-2022.

[39] C. F. Bond Jr and B. M. DePaulo, "Accuracy of deception judgments," *Personality and Social Psychology Review*, vol. 10, no. 5, pp. 214–234, 2006.

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[41] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014, ch. 11, p. 147.