MultiModal Deception Detection: Accuracy, Applicability and Generalizability*

Vibha Belavadi*, Yan Zhou*, Jonathan Z. Bakdash†, Murat Kantarcioglu*,
Daniel C. Krawczyk‡, Linda Nguyen‡, Jelena Rakic‡, Bhavani Thuriasingham*
*Department of Computer Science
University of Texas at Dallas
Richardson, TX 75080
{vxb141530, yan.zhou2, muratk, bxt043000}@utdallas.edu
†U.S. Army Combat Capabilities Development
Command, Army Research Laboratory South at
University of Texas at Dallas
Richardson, TX 75080
jonathan.z.bakdash.civ@mail.mil
†School of Behavioral and Brain Sciences
University of Texas at Dallas
Richardson, TX 75080
{daniel.krawczyk, linda.nguyen, j.rakic}@utdallas.edu

***Abstract*—The increasing use of Artificial Intelligence (AI) systems in face recognition and video processing in recent times creates higher stakes for their application in daily life. Increasingly, critical decisions are being made using these AI systems in application domains such as employment, finance, and crime prevention. These applications are done through the use of more abstract concepts such as emotions, trait evaluations (e.g., trustworthiness), and behavior (e.g., deception). These abstract concepts are learned by the AI system using the verbal and non-verbal cues from the human subject stimuli (e,g., facial expressions, movements, audio, text) for inference. Because the use of AI systems often happens in high stakes scenarios, it is of utmost importance that the AI system participating in the decision-making process is highly reliable and credible. In this paper, we specifically consider the feasibility of using such an AI system for deception detection. We examine if deception can be caught using multimodal aspects such as facial expressions and movements, audio cues, video cues, etc. We experiment using three different datasets with varying degrees of deception to explore the problem of deception detection. We also study state-of-the-art deception detection systems and investigate whether we can extend their algorithm into new datasets. We conclude that there is a lack of reasonable evidence that AI-based deception detection is generalizable over different scenarios of lying (lying deliberately, lying under duress, and lying through half-truths) and that in the future additional factors will need to be considered to make such a claim.**

***Index Terms*—deception detection, multi-modal data analysis, machine learning, ethics, facial expressions**

The research reported herein was supported in part by NIH award 1R01HG006844, NSF awards CICI- 1547324, IIS-1633331, CNS-1837627, OAC-1828467 and ARO award W911NF-17-1-0356. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Combat Capabilities Development Command Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation.

I. Introduction

Are AI/Machine Learning (ML) algorithms accurate and generalizable enough for real-world use? This is the question on many people's minds when they need to interact, directly or indirectly, with an AI/ML-based system for daily decision making. Wrong predictions by these systems may have persistent effects on individuals' lives and society as a whole. For example, AI/ML systems to detect deception have the potential to improve security at border crossings and airport checkpoints [1] but could instead have negative consequences due to limitations in the data set or other factors. Such systems are highly appealing because a quantitative synthesis on human accuracy for detection of deception found it was only slightly above chance on average [2]. Despite these obstacles, airports have long been testing or even implementing AI systems to detect deception in travelers [3], [4]. However, the actual accuracy of such systems in these real-world scenarios are at best questionable [3].

Deception is often considered within the context of criminology. Accordingly, criminology and forensics have been a field with a vested interest in understanding deception on a deeper level. When law enforcement officials – police officers, FBI agents, etc. – interview suspects, they need to be able to detect and differentiate between those who are guilty and lying and those who are innocent and telling the truth. Therefore, most research in deception detection has centered on creating more standard interviewing methods for forensic investigations. Behavioral research in deception detection has traditionally used a variety of stimuli videos including opinion videos and higher stakes crime videos [5], [6].

Systems for detecting deception often rely on facial ex-

pressions, specifically micro-expressions [1]. For decades, facial expressions, specifically micro-expressions, have been considered by some psychologists to be a reliable and valid source of information for detecting deception [7]. Micro-expressions (e.g., a briefly raised eye-brow) are momentary and subtle changes in facial information, often unnoticeable [8], and posited to be useful for lie detection because they are involuntary [7]. Neuroscience suggests facial expressions are controlled by two different brain areas, one for voluntary movement and another one for involuntary movement [9], [10]. The involuntary "leakage" of an expression is proposed to be the clue for deception [11]. Thus, one perspective is that micro-expressions are universal indicators of emotions [12].

Similarly, patterns in eye blinks may also indicate deception [13]. Some studies have found that when people lie, their pupils dilate and the dilation is more when the lie is larger in magnitude [14]. Their eye-tracking data showed that people will look towards the result (in this financial stimulation, the payoff) of the true state disproportionately more than the non-true state. They suggest that if you combine the eye-tracking data and pupil dilation data, the miss rate can be lowered by half. Other research has found that participants were susceptible to cues about "what liars do." It was found that when participants were told liars look to the left, participants were more likely to say that there were more left eye gazes and vice versa for the cue that liars look to the right [15]. Gender has also been found to affect the relationship between eye movement and deception. Some researchers found gender differences in eye-tracking between females and males where females were more likely to make eye contact with one another [16].

However, a contrasting perspective is that facial expressions, including micro-expressions, are not necessarily reliable, valid, or universal indicators of emotion including cues to deception. This is because the meaning of the expression(s) can depend on several factors: the type of emotion, the person, and the specific context (situation and culture) [17]. Thus, facial expressions may be inherently noisy data. Likewise, other neuroscience research indicates there may be many brain systems involved in deception and their patterns of brain activation depending on the type of deception along with the previously mentioned factors [18]. Consequently, there is a risk that computational systems using facial information to detect deception might be overly specific, due to the absence of consistent universal features, and thus lack generalizable accuracy [3].

Recently, multiple papers have evaluated micro-expressions and other features using AI/ML-based techniques sometimes finding high accuracy for detecting lies in video [19]–[25]. Video content can be used to extract facial expressions, body movements, audio, and text. These multi-modal abstract features can be used to train an AI/ML system for learning tasks such as making employment, financial, and security decisions. A highly reliable and valid AI/ML system is desired in such a decision-making process.

In this work, we explore the feasibility of applying AI/ML techniques to detect lies in video using multiple datasets. More importantly, we investigate the accuracy of ML techniques when tested on datasets with varying degrees of deception. A variety of ML algorithms are employed and compared in our study, including random forest, multiple instance learning, and deep neural network. In contrast to some existing work, our empirical results demonstrate the lack of reasonable evidence that ML-based deception detection is generalizable over different scenarios of lying (lying deliberately, lying under duress, and lying through half-truths); and indicate the need for more extensive and in-depth research in evaluating the deception detection claims.

Our main contributions include:

- providing two new video datasets from screening interviews in a mock lying/crime experiment (in each dataset, there are separate videos of the same person being truthful and deceptive);
- deploying a variety of ML techniques to explore the applicability and reliability of these algorithms in deception detection;
- comparing our results to the state-of-the-art ML-based detection systems;
- demonstrating that state-of-the-art ML-based deception detection has a long way to go before it is considered generalizable and reliable.

The rest of the paper is organized as follows. In Section II, we discuss existing work related to our study. In Section III we formally define the problem and present our methodology for deception detection using machine learning techniques. Section IV presents our experimental results and Section V concludes our work.

II. Related Work

Gupta et al. [26] provide a dataset, referred to as Bag-of-Lies, for deception detection with multimodality including video, audio, EEG, and gaze data. The dataset combines vision and the cognitive aspect of deception, collected in a realistic scenario. They also investigated the benefits of incorporating multiple modalities for fusion on the provided dataset.

Rill-García et al. [19] present transfer learning in scenarios where labeled data for deception detection is insufficient. They present a study on the feasibility of using linguistic features for cross-domain deception detection on multimodal data across different domains, from written texts to unrelated topics transcribed from spoken statements. Their empirical study demonstrates the effectiveness of the cross-domain transfer learning technique, reaching an accuracy of 63.64% and an AUC of 0.6351.

In separate work, Rill-García et al. [20] extract features from videos that are highly indicative of deception. Besides, they perform a study of different multimodal fusion methods based on boosting to improve the results obtained by using the different sets of extracted features separately. High-level features are extracted with open automatic tools for the visual, acoustical, and textual modalities. They report experimental results on a real-life trial dataset for deception detection and a Mexican deception detection dataset using Spanish as the spoken language.

Karimi et al. [21] propose an end-to-end framework DEV to detect deceptive videos automatically. Inside the framework, rich and sophisticated features are automatically captured from complicated video data, and interpretable visual cues are extracted from the video. Their technique is robust even when the number of training data is small. Experimental results on real-world video data demonstrate the effectiveness of their proposed framework.

Krishnamurthy et al. [22] train a simple multilayer perceptron (MLP) on multi-modal input. They combine features from different modalities including video, audio, and text along with Micro-Expression features. They reported that their simple model outperforms existing techniques for deception detection on a dataset of real-life deception videos, achieving an accuracy of 96.14% and ROC-AUC of 0.9799.

Ding et al. [23] focus on effectively fusing the useful cues in the face and body for deception detection. For face-body multimodal learning, they propose a novel face-focused cross-stream network (FFCSN) in which face detection is added into the spatial stream to capture the facial expressions explicitly, and correlation learning is performed across the spatial and temporal streams for joint deep feature learning across both face and body. The FFCSN model is trained with both meta-learning and adversarial learning. Their experimental results demonstrate that the FFCSN model achieves state-of-the-art results. They also show that their FFCSN model is generally applicable to other human-centric video analysis tasks such as emotion recognition from user-generated videos.

As a contrast to the papers described above, in this paper, we examine the feasibility of using AI/Machine Learning based deception detection systems. We evaluate our algorithms on three different datasets with varying degrees of deception and show that these automated systems do not perform well on them. We also investigate if there is enough reasonable evidence to show that our AI deception detection system is generalizable and reliable over different scenarios of lying when training and testing on different deception datasets.

III. Methodology

Given a set of videos where truth or lie is known, we define the deception detection problem as follows:

Deception Detection

Input: A set of video input $V = \{v_1, \ldots, v_N\}$ and the corresponding labels $Y = \{y_1, \ldots, y_N\}$ where $y_i \in \{lie, truth\}$.

Question: Is there a machine learning model f that can accurately label v_i for any $v_i \in V$, i.e. $Pr(f(v_i) \neq y_i) < \epsilon$, for an arbitrarily small $\epsilon > 0$.

In all our experiments, we pre-process the video data to form structured input for various machine learning techniques. We first split the video data into training and validation datasets. Next, each video clip in the datasets is disintegrated to create frame by frame data instances, that is, a series of data instances made up of a chosen set of features, as shown in Figure 1. Data instances from the same video clip are assigned the same label, either "lie" or "truth".

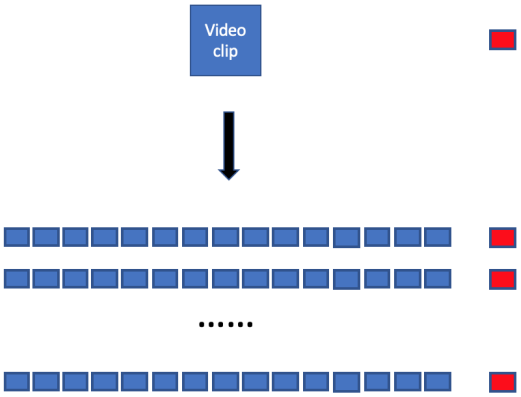

Fig. 1. Breaking a video clip to a series of data instances.

We consider three different sets of features: action units, eye aspect ratio and Convolutional Neural Network (CNN) features, which are discussed in detail next.

A. Action Unit

We used OpenFace [27] to generate frame by frame data for our deception detection problem. Given a video, Openface API returns a CSV file containing rich data regarding the location of the facial landmarks, the confidence of the face recognition algorithm, gaze, head pose tracking, and the action unit data for 18 action units. We are interested in the action unit data extracted by Openface API especially the existence of a particular action unit (represented by AU_c) and the intensity of a given action unit (represented by AU_r). Using these two sets of action units as features, we have different ways to calculate valuation metrics for our experiments using ML models: 1) Per frame-level metric calculation, 2) Per video-level metric calculation using majority voting. In 1) per frame-level metric calculation, we calculate the evaluation metric (accuracy/auc score) by considering all the frames as independent points in the dataset and averaging them out. In 2) per video-level metric calculation using majority voting, we calculate the evaluation metric (accuracy/auc score) by choosing a label for the video using majority voting from the frames of the video.

B. Eye Aspect Ratio

Traditionally, detecting motion in the eye region involves eye detection, optical flow tracking, and heuristic decision on eyelid position [28]–[30]. These approaches are less practical because of their strong sensitivity to the setup and quality of the image/video. Eye Aspect Ratio (EAR) provides a simple and elegant approach to tracking eye dynamics [31]. EAR is computed using the 2D facial landmark locations $p_1, \ldots, p_6$ as illustrated in Figure 2 [31]:

$$EAR = \frac{|p_2 - p_6| - |p_3 - p_5|}{2 \cdot |p_1 - p_4|},$$

which casts the relation between the width and the height represented by the facial coordinates.

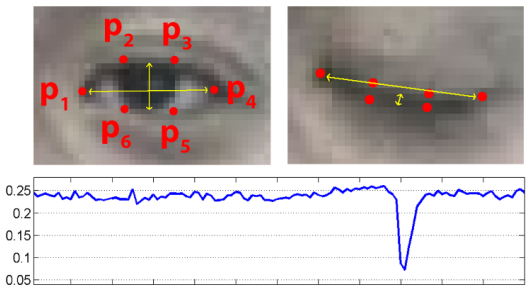

Fig. 2. 2D facial landmarks in the eye region (by Soukupová and Čech [31]).

C. Convolutional Neural Network features

To extract features for our Convolution Neural Network (CNN) model, we ensure that all the videos of a given dataset have the same image dimensions in their video codec. Ensuring the same image dimensions in the video codec will translate to the same image dimensions at the frame-level. We also convert all the videos to 24fps. We build a custom dataset by modifying the standard dataset class of HMDB51 from Pytorch. HMDB51 is an action recognition dataset that extracts video clips of given frame length and frame step between the clips and hence it fits naturally for our use case. For our CNN model, the features are video clips of 12 frames. Each video clip is non-overlapping, i.e, the frame step between the video clips is 12 frames. Considering the frame rate of the videos of 24fps, generating video clips of 12 frames will help capture the micro-expressions of the participants in the dataset.

IV. Experiments

A. Datasets

We considered three different types of deception videos with varying degrees of deception. The purpose of using multiple datasets was to test how much AI/ML models learned for one dataset would be applicable in detecting deception accurately in novel datasets.

The first dataset we use for our experiments is the courtroom trial videos first introduced by [32] collected from public courtroom trials. The authors manually annotated each video with a set of facial expressions, most notable being: Frowning, Eyebrows raising, Lip corners Up, Lips Protruded, and Head Side Turn. The authors of [21] further split a subset of these videos into 4s clips and manually annotated the existence of the above emotions in them. We primarily use these 4s clips and the corresponding annotations in our experiments. We call this the Trial dataset in our experiments. The Trial dataset encapsulates the real-life costs and ramifications of lying in a courtroom scenario.

The second dataset is called the Opinion dataset. In the opinion dataset, participants narrated a movie they watched recently. They were instructed to randomly pick a movie that they liked or disliked and record two narrative videos – one on the movie they liked and the other on the movie they disliked. At the end of the video recording, the participant revealed to the interviewer which was a lie and which was the truth and their video recordings were labeled accordingly. In this dataset, the stakes for lying were considerably lower, and also cover the case of speaking half-truths while lying.

The third dataset in our experiments is the Crime dataset. In the crime dataset, participants were instructed to steal or not to steal 50 U.S. dollars. Both groups had to convince the interviewer that they were innocent. An interviewer who was unaware of which condition the participant was in was sent into the interview room to record their responses. After generic questions, the deception related questions asked were 1) “Did you steal the money?” 2) “Are you lying to me now?”. This dataset captures higher stakes of lying compared to the opinion dataset, but less than that of the trial dataset. This data set focuses on lying as it more closely appears in daily life.

We ran experiments on all three datasets and tried to generalize them by training on one dataset (e.g., Opinion) and transferring it to another (e.g., Crime and Trial). Specifically, we consider training on the Opinion dataset and testing on the Trial and Crime datasets. Since the Crime dataset has different sections of the video where a person can be lying or telling the truth, we focus on only the specific portion of the video where the person is either truthful or lying: 1) "Did you steal the money" and 2) "Are you lying to me now?". We extract our features between the two starting and ending snapshots and use that for testing our Opinion dataset model.

B. Random Forest Method

Our Random Forest method considers the the deception classification using action units on two variants of the dataset: per frame-level and 4s video-level. Since Openface data applies the action units in two different flavors: 1) existence of action unit in a video and 2) intensity of action unit in the video, we try four different combinations on the frame-level: 1) AU_exist_in_video, 2) AU_exist_in_video_norm, 3) AU_intensity_in_video and 4) AU_intensity_in_video_norm. To get 2) and 4) from 1) and 3) we sum up the values in 1) and 3) and then normalize them. We measure the best validation accuracy and the corresponding training accuracy for each of the four scenarios. To ensure stable results and reduced bias, we have also calculated 10 fold cross-validation. Tables I and II summarizes the results for Random forests run on frame-level for the datasets. In both the tables, we can see that it overfits the training data but fails to generalize on the validation set. Also, the Crime dataset has better results compared to the Opinion dataset in terms of validation accuracy. This might be because of the differences in the dataset content and the different contexts of lying in both the datasets. We can see that normalisation (AU_exist_in_video_norm and AU_intensity_in_video_norm) results in a drop in validation accuracy compare to the non-normalized counterparts (AU_exist_in_video and AU_intensity_in_video)

In addition to the frame-level experiments, we also ran the Random Forest algorithm on the 4s video clips. We extracted all the frames corresponding to the 4s clip of the Crime, Opinion, and Trial dataset and assigned the AU_c (Action

TABLE I
10-FOLD CROSS VALIDATION ACCURACY OF RANDOM FOREST ON THE OPINION DATASET: FRAME-LEVEL

model	kfold-train	kfold-validation
AU_exist_in_video	63%	48%
AU_exist_in_video_norm	77%	38%
AU_intensity_in_video	80%	38%
AU_intensity_in_video_norm	78%	38%

TABLE II
10-FOLD CROSS VALIDATION ACCURACY OF RANDOM FOREST ON CRIME DATASET: FRAME-LEVEL

model	kfold-train	kfold-validation
AU_exist_in_video	58%	63%
AU_exist_in_video_norm	67%	55%
AU_intensity_in_video	67%	62%
AU_intensity_in_video_norm	58%	59%

TABLE III
10-FOLD CROSS-VALIDATION ACCURACY OF RANDOM FOREST ON ALL DATASETS: 4S CLIP-LEVEL

model	overall_acc_score	acc_score_maj_label	acc_video
Crime_4s	46%	38%	44.2%
Opinion_4s	32%	28%	33.4%
Trial_4s	53%	59%	57.8%
Op_on_Cr_4s	50.6%	43.3%	48.5%

TABLE IV
10-FOLD CROSS-VALIDATION AUC-SCORE OF RANDOM FOREST ON ALL DATASETS: 4S CLIP-LEVEL

model	overall_auc_score	auc_score_maj_label
Crime_4s	0.42	0.40
Opinion_4s	0.24	0.28
Trial_4s	0.58	0.56
Opinion_on_Crime_4s	0.48	0.43

Unit existence in the frame) column value as one if any of the individual frames of the 4s video clip had the action unit and zero otherwise. We only chose Action unit existence for our experiments on 4s clips. For the Trial dataset, we did not use the Openface action units, but rather the manual annotations of micro-expressions (head side turn, lips protruded, etc.) provided by the authors of the dataset [32]. We also used 10-fold cross-validation to get more stable results. The results of our experiments are shown in tables III and IV, which measure the accuracy and auc score metrics for the validation set respectively.

As is consistent with the frame-level results, Random Forest performs better on the Crime dataset compare to the Opinion dataset. The Trial dataset has the best performance among the three datasets, going slightly above random on the validation set. This might be attributed to the more handcrafted features for the dataset (head side turn, lips protruded, etc.). Unfortunately, the accuracy isn't sufficiently large to claim a reliable model. We also tested the transferability of our model (trained on opinion dataset) on the Crime dataset as shown in the last row of tables III and IV ('Op_on_Cr_4s model'), and we can conclude that the accuracy and the auc score of detecting deception on Crime dataset marginally improves when a model trained on Opinion dataset is used on it. However, it is not significant enough to warrant a generalization that these deception models help. From the experiments, it is clear that given micro-expression intensity or existence in the video, it is not possible to confidently predict deception. Also, the deception detection models are highly specific to their dataset and cannot be transferred to test other datasets that vary greatly from the original dataset.

C. Multiple Instance Learning Method

Multiple Instance Learning [33] is employed in the case of an ambiguous relationship between whole-part. In our problem formulation, we put all the frames of a video in a single bag. Since it is a binary classification multiple instance learning, we label a video as negative if all the frames in the video bag are negative. If there is at least one positive label frame in the video bag, the video bag is labeled as positive. We use Action Unit intensity as our features for the model and provide our results on the 10-fold validation accuracy. We try three different variants of multiple instances learning: MissSVM, sbMIL, and SIL on the opinion dataset. Table V shows the results for these three different approaches of Multiple Instance Learning.

In the Single Instance Learning (SIL) method, each instance (frame in our case) is assigned the label of the bag (video). sbMIL or sparse balanced Multiple Instance Learning has a hyperparameter η that represents a fraction of positive examples in the bag where the top η instances in the bag are labeled positive while the rest are labeled negative. Finally, in the case of MissSVM, a standard SVM is employed for the Multiple Instance classification.

The results in the table are extremely poor and barely reaching 50% in the SIL variant. A major drawback of multiple instance learning is that it has no guaranteed optimal solution since it uses local optimization heuristics [34].

TABLE V
ACCURACY OF MULTIPLE INSTANCE LEARNING ON OPINION DATASET

model	rbf_kernel	quadratic_kernel	linear_kernel
MissSVM	50.0%	14.3%	25.0%
sbMIL	3.6%	10.7%	14.3%
SIL	50.0%	50.0%	50.0%

D. CNN Method

We applied the CNN method for deception detection using video frames extracted with ffmpeg [35]. We used the pre-trained ResNet-3D (r3d_18) model from Pytorch as a base for our experiments. Resnet 3D model is an 18 layer video classification model [36] that captures spatio-temporal features through a space-time convolution block ((2+1)RD). We choose an action recognition model as a base since we are trying to calculate the action of 'lying' or 'telling the truth'. The

model takes as input a series of 12s frame videos with the ground truth of the original video. We experimented with two different transfer learning techniques: using as a feature extractor or finetuning the model, and found out that finetuning gave better accuracy. We used 10-fold cross-validation to record the accuracy of different variants. For the three variants of the accuracy: the overall validation accuracy (val_acc), validation accuracy using majority label (val_acc_maj_label) and validation accuracy averaged using per video accuracy of the validation set (val_acc_avg_video), it is seen that Crime dataset has better accuracy compared to the CNN model trained on opinion for the overall validation accuracy. However, despite the improvement of the other two metrics based on average per video accuracy and majority label, when the model trained on the Opinion dataset is used, none of the prediction accuracy results are above random guessing.

TABLE VI
ACCURACY OF CNN ON OPINION AND CRIME DATASET

dataset	val_acc	val_acc_maj_label	val_acc_avg_video
opinion	44.49%	46.26%	45.95%
crime	54.85%	43.57%	44.17%
trial	50.19%	46.82%	46.92%
op_on_crime	44.49%	48.12%	48.15%

E. Learning on Eye Dynamics

In this experiment, we narrow down our study to a key facial landmark dynamic—the movement of eyes, i.e. eye opening and closing in each video frame. Figure 3 illustrates an example of eye dynamics measured in EAR in a "lie" video and a "truth" video in the *Crime* dataset. Over the entire dataset, there is an indication that liars are relatively more inclined to have consistent eye dynamic patterns while truth-tellers are more relaxed and often register a larger range of eye movement. It has been reported that people tend to blink eight times more frequently right after they lie [37].

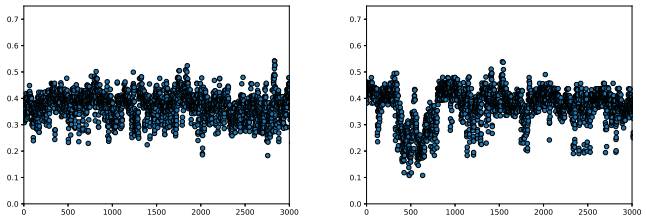

Fig. 3. EAR data sequences of a *lie* and a *truth* video.

We train a Deep Neural Network (DNN) model on the EAR data of the *Opinion* dataset and test on the *Crime* and the *Trial* datasets. The *Opinion* dataset is used as the training data because it consists of individuals that are either lying or telling the truth throughout the clip, therefore the labels of data sequences created from each clip are accurate. The DNN model consists of an input layer, two fully connected layers with 50 units each, and an output layer that predict either `lie` or `truth`. We use the `tanh` function as the activation function on the hidden layers.

Each video is pre-processed using the sliding window approach to generate a series of EAR data sequences with 1000 frames per sequence, as shown in Figure 1. Each data sequence is created by sliding the window 100 frames down stream in the video clip. Data sequences from the same video clip have the same label, either "lie" or "truth".

When making predictions for the *Crime* data, we aggregate the predictions for data sequences from the same video clip and predict the entire clip as "lie" if the aggregated result is greater than a threshold. The threshold value is selected to maximize the training accuracy with desired false positive rates.

We report the training and cross-validation accuracy on the *Opinion* dataset, and the test accuracy on the *Crime* and the *Trial* datasets. Note that the *Opinion* dataset and the *Crime* dataset are video clips taken from the same group of individuals, while the *Trial* dataset comes from a different research group. All experiments were run 10 times and the averaged results were reported. Figure 4 shows the accuracy on each dataset, error bars are one standard error of the mean. As can be observed,

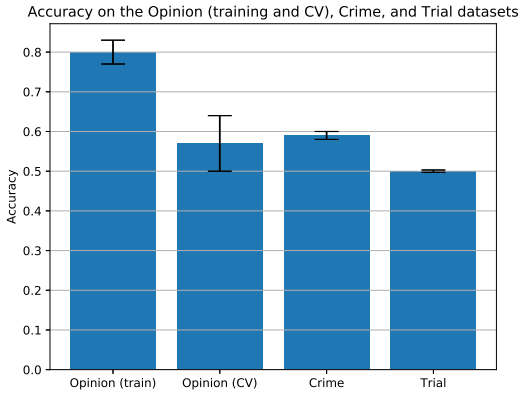

Fig. 4. Accuracy on the three datasets.

the DNN model well trained on the *Opinion* dataset achieved much better results (approximately 60%, with an average false positive rate of 0.083) on the datasets collected from the same individuals, but failed to make reliable and meaningful predictions (approximately 50%) for other deception datasets created from different groups in a different context.

F. Histograms of Eye Blinks

In this experiment, we compare the histograms of the percentage of blinks in the entire video for the lie and truth videos, as shown in Figure 5. Clearly, there are more people who blink less (blink percentage $\leq$ 2%) when they lie, and there are more people who blink more (blink percentage $\geq$ 4%) when they tell truth. However, there is no such a clear-cut threshold or general probabilistic pattern that separates the lying and the truth-telling groups. However, when we pair the truth/lie videos of the same individual, we can see that the majority of the truth videos have higher blink percentage than

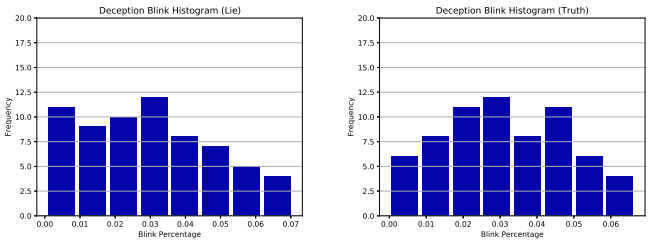

Fig. 5. Percentage of blinks during “Lie” and “Truth” clips.

their lie counterparts. Figure 6 illustrates the truth-to-lie ratio of blinks of video clips from the same individuals who are telling the truth or lying. It is clear the same individual tend to blink more when not lying (truth-to-lie ratio > 1).

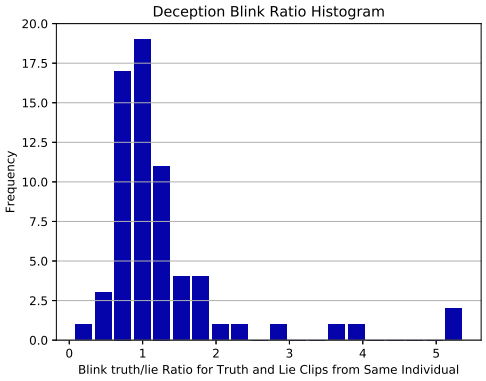

Fig. 6. Same individual truth/lie blink ratio.

We take the *Crime* data, and compute the percentage of blinks p_i in each video of the individual with an ID i. Figure 7 shows the histograms of the “lie” and “truth” videos in the *Crime* data. Suppose we know the percentage of blinks p_i^ℓ for the lying individual that has the same ID in the *Opinion* dataset, we compute the ratio $r = \frac{p_i}{p_i^\ell}$. If $r > 1$, we predict the crime video as “truth”, otherwise, it is a “lie”. The accuracy of this simple detection technique is 56.2%. This level of accuracy is similar to human detection of deception [2].

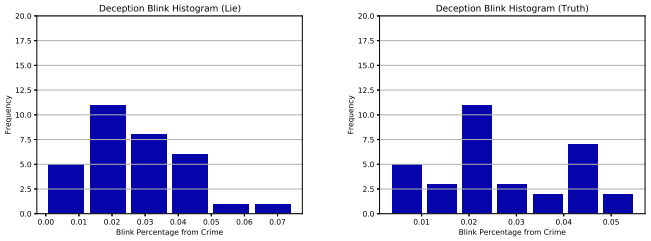

Fig. 7. Percentage of blinks during “Lie” and “Truth” clips on Crime.

V. Conclusion

Our findings indicate, that for a variety of features and models, using multimodal data for detection of deception does not appear to be generalizable. Transfer performance was around chance or even less. This shows there is insufficient evidence that AI systems for detecting deception are likely to achieve adequate accuracy in real-world use. This observation may seem like a contraction given the previous work that achieves high deception detection accuracy on certain datasets. Our results suggest that such results are more likely to occur due to well chosen features that overfit to the given dataset, and they may not generalize to other datasets. For example, using slightly different features, we could not reach the accuracy reported for the original Trial dataset [32]. Please note that in [32], authors coded "facial features" that they define and used these important features in detecting the deceptive behavior.

Our findings are consistent with psychological research that facial expressions and other information may not be universal probabilistic clues to deception. Instead, the information in facial expressions may be highly dependent on numerous factors including the person, situation, and culture [17]. Here, we were able to evaluate the person in two datasets, the same participants told the truth and a lie in separate videos. In addition, we also evaluated distinct situations with varying levels of deception using three datasets. However, we did not evaluate potential cultural differences in deception.

A. Ethics

Recent work has stated there are "... unclear definitive and reliable cues to deception, we question the validity of using artificial intelligence that includes cues to deception, which have no current empirical support" [3], p. 1. We agree with this statement —our results are consistent with the lack of empirical support for reliable AI detection of deception. This is particularly concerning because AI systems for deception detection are being used and tested in the real-world. More generally, the use of potentially flawed AI systems in the real-world could have negative societal impacts and this also raises ethical concerns [38], [39]. Therefore, further cross cultural studies using diverse datasets are needed to evaluate the accuracy of the AI/ML based deception detection.

References

[1] T. O. Meservy, M. L. Jensen, J. Kruse, J. K. Burgoon, J. F. Nunamaker, D. P. Twitchell, G. Tsechpenakis, and D. N. Metaxas, “Deception detection through automatic, unobtrusive analysis of nonverbal behavior,” *IEEE Intelligent Systems*, vol. 20, no. 5, pp. 36–43, 2005.

[2] C. F. Bond Jr and B. M. DePaulo, “Accuracy of deception judgments,” *Personality and Social Psychology Review*, vol. 10, no. 3, pp. 214–234, 2006.

[3] L. M. Jupe and D. A. Keatley, “Airport artificial intelligence can detect deception: or am i lying?” *Security Journal*, pp. 1–14, 2019.

[4] S. Weinberger, “Intent to deceive?” *Nature*, vol. 465, no. 7297, pp. 412–416, 2010.

[5] M. Frank and P. Eckman, “Appearing truthful generalizes across different deception situations,” *Journal of Personality and Social Psychology*, vol. 86, 2004.

[6] M.-A. Reinhard, R. Greifeneder, and M. Scharmach, “Unconscious processes improve lie detection,” *Journal of Personality and Social Psychology*, vol. 105, 2013.

[7] P. Ekman, “Lie catching and microexpressions,” *The Philosophy of Deception*, vol. 1, no. 2, p. 5, 2009.

[8] D. Matsumoto and H. C. Hwang, "Microexpressions differentiate truths from lies about future malicious intent," *Frontiers in Psychology*, vol. 9, p. 2545, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2018.02545

[9] W. Rinn, "The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions," *Psychological bulletin*, vol. 95, pp. 52–77, 02 1984.

[10] D. Matsumoto and M. Lee, "Consciousness, volition, and the neuropsychology of facial expressions of emotion," *Consciousness and Cognition*, vol. 2, no. 3, pp. 237–54, 1993.

[11] P. EKMAN, "Mistakes when deceiving," *Annals of the New York Academy of Sciences*, vol. 364, pp. 269 – 278, 12 2006.

[12] P. Ekman, "Facial expressions of emotion: an old controversy and new findings," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 335, no. 1273, pp. 63–69, 1992.

[13] K. Fukuda, "Eye blinks: new indices for the detection of deception," *International Journal of Psychophysiology*, vol. 40, no. 3, pp. 239–245, 2001.

[14] J. T. yi Wang, M. Spezio, and C. F. Camerer, "Pinocchio's pupil: Using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games," *American Economic Review*, vol. 100, 2010.

[15] F. Spiroiu, "The impact of beliefs concerning deception on perceptions of nonverbal behavior: Implications for neuro-linguistic programming-based lie detection," *Journal of Police and Criminal Psychology*, 2018.

[16] M. Zhang, T. Liu, M. Pelowski, and D. Yu, "Gender difference in spontaneous deception: A hyperscanning study using functional near-infrared spectroscopy," *Scientific Reports*, vol. 7, 2016.

[17] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 1–68, 2019.

[18] K. E. Sip, A. Roepstorff, W. McGregor, and C. D. Frith, "Detecting deception: the scope and limits," *Trends in Cognitive Sciences*, vol. 12, no. 2, pp. 48–53, 2008.

[19] R. Rill-García, L. Villaseñor-Pineda, V. Reyes-Meza, and H. J. Escalante, "From text to speech: A multimodal cross-domain approach for deception detection," in *Pattern Recognition and Information Forensics*, Z. Zhang, D. Suter, Y. Tian, A. Branzan Albu, N. Sidère, and H. Jair Escalante, Eds. Cham: Springer International Publishing, 2019, pp. 164–177.

[20] R. Rill-Garcia, H. Jair Escalante, L. Villasenor-Pineda, and V. Reyes-Meza, "High-level features for multimodal deception detection in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[21] H. Karimi, J. Tang, and Y. Li, "Toward end-to-end deception detection in videos," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 1278–1283.

[22] G. Krishnamurthy, N. Majumder, S. Poria, and E. Cambria, "A Deep Learning Approach for Multimodal Deception Detection," *arXiv e-prints*, p. arXiv:1803.00344, Mar. 2018.

[23] M. Ding, A. Zhao, Z. Lu, T. Xiang, and J.-R. Wen, "Face-focused cross-stream network for deception detection in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[24] Z. Wu, B. Singh, L. S. Davis, and V. S. Subrahmanian, "Deception detection in videos," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, S. A. McIlraith and K. Q. Weinberger, Eds. AAAI Press, 2018, pp. 1695–1702. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16926

[25] A. C. Elkins, N. Sorros, S. Zafeiririou, J. K. Burgoon, M. Pantic, and J. F. Nunamaker, "Do liars blink differently? automated blink detection during deceptive interviews," in *Proceedings of the Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium, Matthew Jensen, Thomas Meservy, Judee Burgoon, and Jay Nunamaker (Eds.). Hawaii International Conference on System Sciences, Hawaii*, 2014.

[26] V. Gupta, M. Agarwal, M. Arora, T. Chakraborty, R. Singh, and M. Vatsa, "Bag-of-lies: A multimodal dataset for deception detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[27] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 2018, pp. 59–66.

[28] M. Divjak and H. Bischof, "Eye blink based fatigue detection for prevention of computer visionsyndrome," in *IAPR Conference on Machine VisionApplications*, 2009.

[29] T. Drutarovsky and A. Fogelton, "Eye blink detec-tion using variance of motion vectors," in *ComputerVision - ECCV Workshops*, 2014.

[30] W. H. Lee, E. C. Lee, and K. E. Park, "Blink detec-tion robust to various facial poses," *Journal of Neu-roscience Methods*, Nov. 2010.

[31] T. Soukupová and J. Cech, "Real-time eye blink detection using facial landmarks," in *21st Computer Vision Winter Workshop*, 2016.

[32] 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, pp. 59–66.

[33] G. Doran and S. Ray, "A theoretical and empirical analysis of support vector machine methods for multiple-instance classification," *Machine Learning*, vol. 97, 10 2014.

[34] D. F. ul Amir Afsar M. Multiple instance learning. [Online]. Available: https://piazza.com/class_profile/get_resource/ij10b396m3315h/in7tpecafge3jg

[35] ffmpeg. ffmeg. [Online]. Available: https://ffmpeg.org/

[36] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[37] S. Leal and A. Vrij, "Blinking during and after lying," *Journal of Nonverbal Behavior*, vol. 32, no. 4, pp. 187–194, 2008.

[38] K. Crawford and R. Calo, "There is a blind spot in ai research," *Nature*, vol. 538, no. 7625, pp. 311–313, 2016.

[39] ProPublica, "Machine bias," https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016.