Unmasking Lies: A Literature Review on Facial Expressions and Machine Learning for Deception Detection

Monica Sen, Rébecca Deneckère

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2024) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024)

Unmasking Lies: A Literature Review on Facial Expressions and Machine Learning for Deception Detection

Monica Sen, Rébecca Deneckère

Centre de Recherche en Informatique, University Paris 1 Panthéon-Sorbonne, Paris, France

Abstract

Is there observable behavior or evidence capable of distinguishing a liar from a truthful person? The question has intrigued for centuries, reflecting the inseparable nature of lying in our social interactions. Faced with the human inability to discern truth from falsehood and the limitations of polygraphs, the emergence of facial recognition technologies appears as a promising alternative. Although the use of facial expressions, including micro and macro-expressions, in lie detection is subject to debate, some researchers argue that these nonverbal cues could be revealing. This state-of-the-art review explores facial expression analysis through machine learning for lie detection, questioning the quality of data required to develop effective models.

Keywords: Literature review; Facial expression; Machine learning; Deception; Lie

1. Introduction

Most researchers in the deception detection field agree that there is no "Pinocchio's nose" that can serve as a reliable indicator to easily spot a lie [1]. Deception is a pervasive phenomenon in our societies, occurring multiple times daily [2]. Despite its frequency, humans' ability to detect it seems no better than chance. On average, individuals correctly identify a lie only 47% of the time [3]. Surprisingly, adults do not fare better in discerning lies in children, with a success rate of 47.5% [4]. Although figures may vary across studies, they generally hover around 50%, akin to flipping a coin. Among traditionally used methods, the polygraph is the most widespread. However, its intrusive nature requiring physical connection to the individual's body during interrogation [5] and necessitating the suspect's consent, often reluctant to undergo the test [6], raises concerns. Aware of surveillance, individuals may strategize to outsmart the device. With proper training, suspects can feign innocence using specific techniques like lying during baseline questions, tensing muscles, or biting their tongue [7]. Sensors can also affect the suspect's psychological stability, complicating lie detection [8]. The polygraph has frequently proven fallible [9],

incriminating innocents while exonerating guilty parties. Moreover, the risk of bias inherent in human intervention needed for testing [3], coupled with reliance on physiological responses like blood pressure, heart rate, skin conductivity, muscle tremors, and breathing during interrogations [10], renders its large-scale application unfeasible.

The advancement of artificial intelligence technologies marks a turning point in our ability to understand and analyze human behavior [1]. While AI aims to mimic the human brain, understanding human behavior proves challenging given the inherently complex nature of behavioral science. Human behavior emerges from an interaction between three fundamental elements: actions, cognition, and emotions. Lying is identified as a cognitive behavior [7]. Cognition, as explored in philosophy and political psychology, refers to the conscious and intentional processes underlying thought and knowledge such as perception, attention, memory, language, learning, reasoning, judgment, and higher-order thinking that can be deliberately controlled [11]. The foundation of human psychology is shaped through reactions to emotion-driven behaviors. Emotions such as joy, fear, sadness, anger, surprise, excitement, guilt, regret, hatred, and curiosity, although distinct from behaviors, play a role in motivating certain actions [7].

In the field of mental health research, AI offers an approach that addresses issues from a statistical perspective to optimize prediction. Initially, unimodal channels such as audio, images, text, and physiological channels like ECG, EEG, GSR, BVP, etc., were used to predict and classify cognitive behavior using AI/ML models. Later, multimodal models demonstrated strong multimodal understanding and generalization abilities across several downstream cognitive tasks. Multimodally trained visual and linguistic encoders, which align more closely with the brain than unimodally trained ones from a neuronal encoding perspective, serve as resources for neuroscientists [12]. With access to vast amounts of data, machine learning proves particularly suited to address complex challenges, such as anticipating individuals at risk of mental disorders, predicting emotions, identifying intentions, and detecting lies [7].

Deception has been extensively researched in psychology, leading to numerous studies on detecting deceptive behavior in applied technologies, involving linguistic, behavioral, and physiological domains [13]. Lie detection techniques can use non-verbal characteristics. They include facial expressions but can encompass other signs such as eye or mouth movements, response time, swallowing frequency, or facial symmetry [14][15][16]. Facial Expressions are an aspect of human behavior recognized as the most salient and influential aspect of human communication. The term "facial expression" is used by researchers to define certain recurring movements of facial muscles that convey thoughts, emotions, or behaviors [17]. They can be categorized into two types: macro-expressions and micro-expressions. It is the duration of the facial expression, not its intensity, that differentiates macro-expressions from micro-expressions. Macro-expressions last between 0.75 and 2 seconds and are easily perceived by humans, while micro-expressions, being briefer, last less than 0.5 seconds [18]. Technically, the term Action Unit (AU) corresponds to the fundamental actions of individual muscles or groups of muscles that are activated during facial expressions. Each micro-expression is generated by a specific pattern of AU activation. [19] The Facial Action Coding System (FACS), developed by Paul Ekman and Wallace V. Friesen, is a coding system aimed at describing all possible facial expressions using action units. Facial movements were originally classified manually, resulting in the definition of 46 AUs to identify up to 7,000 different AU combinations [19]. This nomenclature has been widely used in lie detection research to identify differences in AU activation between truthful and deceptive statements [20].

These last years have seen several proposals to detect lies from facial expressions by using machine learning techniques. But how do these techniques work, and do they succeed to detect deception? This paper tries to answer this question by exploring the literature with a systematic review of this research field.

This paper is structured as follows. The first section explains the research method. Section 3 proposes the state of the art of the research question and we conclude in section 4.

2. Research Method

In this section, we present the research methodology used for the systematic literature review. The research method used followed the guidelines presented by [21] which is composed of three main steps: planning, conducting, and reporting the review.

Planning the review. Given the limitations associated with polygraph use and the need to identify deception in various contexts without human intervention, researchers are turning to machine learning-based solutions. Humans are accustomed to observing and understanding cues by looking at them, so the most natural approach to analyze these signs would be to record videos or take photos. In this context, the issue guiding this thesis is **Can machine**

learning help detect deception from facial cues? Note that the term "deception" is not strictly limited to expressing false things but includes deceptive behaviors such as omission, pretense, and any other action aimed at misleading. In the deception detection research field, verbal and non-verbal cues are distinguished but this work is restricted to non-verbal cues, especially facial expressions.

Conducting the review. This step aims at identifying a set of papers. We first selected them based on a relevant search string in the SCOPUS scientific database using the SCOPUS Search API. We used a set of keywords like detection, lie, lying, expression, facial expression, face recognition, etc., mixed with the terms machine learning or deep learning. We obtained 108 papers. We then selected our corpus of papers following several inclusion and exclusion criteria to finally obtain 8 papers: [14][15][20][22][23][24][25][26].

Inclusion criteria: The paper is a result of the search string (TITLE-ABS-KEY("detection") OR TITLE-ABS-KEY("detect") OR TITLE-ABS-KEY("detector") OR TITLE-ABS-KEY("detecting")) AND (TITLE-ABS-KEY("lie") OR TITLE-ABS-KEY("lying") OR TITLE-ABS-KEY("liar") OR TITLE-ABS-KEY("deception")) AND (TITLE-ABS-KEY("micro-expression") OR TITLE-ABS-KEY("facial expression") OR TITLE-ABS-KEY("face recognition") OR TITLE-ABS-KEY("facial analysis")) AND (TITLE-ABS-KEY("machine learning") OR TITLE-ABS-KEY("deep learning")); The paper addresses at least one of the following topics: Facial expression recognition, Emotion recognition, Multimodality including facial expressions, specifically in of deception detection using machine learning techniques.

Exclusion criteria: The paper is not in English; It is not available online; The paper doesn't report an experiment.

Reporting the review. A generic process for detecting lies has been identified in the selected papers. Each of the process steps has been studied in all the papers of the corpus and explained in the analysis section.

Validity Threats. we studied five possible threats, which could affect the validity of our research [27]. To lessen the validity threats that could impact our study, we present them with suitable modification actions in Table 1.

Table 1. Validity Threats.

Validity	Actions
Descriptive validity	We unified the concepts and criteria used in the study and structured the information to be collected with a data extraction form to support a uniform recording of data.
Theoretical validity	We used a search string and applied it to SCOPUS API, which includes the most popular digital libraries on computer sciences and software engineering. A set of inclusion and exclusion criteria has been defined. We used only an automatic search, without snow-bowling strategy, thus we increased the risk of not finding all the available evidence. The choice of English sources should be of minimal impact.
Generalization validity	Our research question is general enough to identify and classify the findings.
Evaluative validity	Two researchers studied the papers, working independently but with an overlap of studies to identify potential analysis differences. Two researchers validated every conclusion.
Transparency validity	The research process protocol is detailed enough to ensure it can be exhaustively repeated.

3. Machine Learning for Deception Detection based on Facial Expressions

All articles in the corpus propose studying lie detection through video analysis and the use of a prediction model. The steps can be summarized as illustrated in Fig. 1.

Fig. 1. Proposals experimental process.

1.Creation of dataset: The ML/DL prediction model is trained on an existing database or a dataset. An experiment is conducted in the laboratory to replicate situations where the subject may lie or exhibit deceptive behaviors. Subjects are filmed during the experiment, creating a set of videos on which the model can work.
2.Extraction of Cues: They are extracted from the videos/images using algorithms or models for lie prediction.
3.Classification of the results: The prediction model is applied to classify videos based on the presence of deception.
4.Model Evaluation: The model's performance in predicting deception is analyzed using various statistical metrics.

3.1. Creation of dataset

This section aims to synthesize the experimental protocols to describe the scenarios and mechanisms used to study deception. This also sets the context for analyzing deception cues. The cues come from either a set of public datasets, in which case researchers [14][23] focus on the model's ability to predict deception, or from a dataset created through an experiment. For clarity, the names of the different datasets used will be emphasized. These methods can be summarized into 3 categories, reflecting the level of deception involved: High, Moderate, and Low-stakes situations.

High-Stakes Situations. High-stakes situations refer to contexts where the consequences of being caught lying are significant. None of the studies in the corpus replicated a situation with high stakes. However, [14][26] used the Real-life Trial dataset, which contains videos of courtroom trials where the verdict determines whether a statement is truthful or deceptive. [14] uses the DDCIT dataset, uses polygraphic techniques to measure the subject's Galvanic Skin Response (GSR) in an attempt to simulate an environment similar to criminal interrogations. The subject being aware that their physiological reactions are being recorded can create a sense of immersion that may heighten the perception of stakes in the experience. Despite the presence of a team of professionals in the field of deception detection, the situations created in the DDCIT dataset cannot be compared in stakes to real cases of offenses and crimes. In [24], people were recorded in a game show with very personal questions, which can be considered as a high-stake situation.

Moderate-Stakes Situations. They involve contexts where the consequences of lying are moderate, often related to common social or professional interactions. Methods include mock crimes, card games, and interrogations. In [26], participants are asked to steal or not steal $50 and then convince the experimenter of their innocence. In [25], a child is prompted to lie by omission about toys he allegedly broke under the threat of getting into trouble. In [14][22], a financial reward is given to participants if they successfully deceive the experimenter, which helps stimulate participants' motivation to be persuasive and continue participating in the experiment. An average score of 6/7 in [22] indicates that participants understood the instructions well and desired to win, resulting in more authentic reactions. The experiment conducted in [22] allows replicating a situation where participants can choose their gameplay strategy and consequently the type of lie they will use. In research, simple deception is often studied, neglecting a more subtle type of deception known as sophisticated deception prevalent in competitive contexts such as political rivalry, war, sports, gambling (like poker), business, and diplomacy, aimed at misleading others.

Low-Stakes Situations. Low-stakes situations represent contexts where the repercussions of lying are minimal. Methods used by [15][20][26] include role-playing and storytelling, where participants are assigned a role as either a truth-teller or a liar and answer open-ended questions on mundane topics. In [26], participants give their opinions on a randomly chosen movie (Opinion dataset). In [20], participants are interviewed about past vacations, whether fictional or real (Holiday dataset). The use of holiday memories as a lying subject has been adopted in previous research [28], as remembering vacations engages the same cognitive processes as providing an alibi during a criminal investigation but in a low-risk context. In the experiments conducted in [20][26], participants were free to express themselves for a set duration, while in [15], participants were restricted to answering only with "Yes" or "No."

Datasets. Parameters such as perceived stakes in a given situation and the lying strategy adopted by the subject must be considered to contextualize the results obtained from ML model predictions. Laboratory experiments may not always reflect real-world lying contexts. If there are facial markers that differ between lying strategies, confusing them could lead to incorrect interpretations. Table 2 synthesizes the datasets of the corpus. The “labelling” field corresponds to how the videos are labeled (truth/lie), helping understand how objective truth is known. For Bag of Lies, Crime, Opinion, and Holiday, participants reveal to the experimenter whether they chose to lie or not. For Mood, instructions to be truthful, lie partially, or lie completely are given each question session. For Joker and DDCIT, truth is obtained when the experimenter checks the participant's cards, which cannot be disputed. For Real-life Trial, videos are labeled based on the judicial verdict: guilty is labeled as a lie, not guilty and acquitted are labeled as truth. The stakes level is not provided in each paper but is arbitrarily determined based on the provided information.

3.2. Extraction of cues (Lie Indicators)

There are various indicators, in facial expressions, that can be used to detect lies: micro and macro expressions, of course, but also emotions directly expressed on the face. There are approaches that use hybrid methods by combining different indicators altogether.

Table 2. Datasets.

Dataset	Real life trial	Bag of lies	Crime	Opinion	Joker	Holiday	DDCIT	Mood	Toys	Personal
Papers	[14][26]	[23]	[26]	[26]	[22]	[20]	[14]	[15]	[25]	[24]
Participants number	56	35	-	-	40	62	105	15	158	16
Participants details	21 females 35 males age: 16 - 60	10 females 25 males			20 females 20 males age: 18 - 23	43 females 19 males age: 20 - 29	51 females 54 males age: 20 - 30	7 females 8 males age: 25 - 33	7 girls 8 boys age: 4 - 9	8 females 8 males
Origin	-	-	-	-	Chinese	Italian	Korean	Korean	Latino, African, American	Chinese
Method	Real life judicial trial	Telling a story based on photos	False crime (committed or not) - thief of 50$	Free speech (movie opinion)	Cards deck (Joker)	Free speech (holidays)	Question on an hidden information (yes/no)	Open question (yes/no)	False crime (not committed) - broken toys	Personal Questions
Stakes level	High	Low	Moderate	Low	Moderate	Low	Moderate	Low	Moderate	High
Labelling	Judicial verdict	participants	participants	participants	cards	participants	cards	Given instructions		polygraph
Data size	121 videos ▪61 lies ▪60 truth	325 videos ▪162 lies ▪163 truth	-	-	120 images 240 videos ▪32 lies ▪30 truth	62 videos ▪32 lies ▪30 truth	630 videos ▪210 lies ▪420 truth	-	High proportion on negative classes	32 videos ▪16 lies ▪16 truth
Existing Dataset	yes	yes	no	no	no	no	no	no	no	no

Macro and Micro-Expressions. The use of facial expressions, particularly micro-expressions, is based on the theory of involuntary expression "leakage". Micro-expressions are believed to be universal indicators of emotions and could reveal our true intentions [19]. Studies in the corpus [14][20][22][23][26] use the OpenFace tool to extract facial expressions. The results obtained allow measuring the presence of certain Action Units and/or evaluating the intensity of an AU. For example, in [14], occurrences of micro-expressions are counted based on the duration of AUs. Considering that a micro-expression lasts up to 0.5 seconds, AUs that appear for less than this limit are counted. Expressions that appear in only one frame are considered detection errors and excluded. In [23], the obtained features are not used as is. The data undergo pre-processing through several steps: Apex Frame Selection (selecting the most expressive images from a video sequence); Feature Selection (choosing the most relevant features for classification. In the study, the most relevant AUs for identifying lies are AU14 (lip corner depressor), AU23 (lip tightening), and AU12 (lip corner puller)); Data Discretization (transforming continuous features into discrete values, which helps reduce the effects of outliers and simplifies the data structure); and Feature Scaling (normalizing feature values ensures that all features contribute equally to the learning process). For [26], preprocessing of data is also necessary to train the CNN model. Researchers adjusted all videos to have the same image size and frame rate (24 fps) and used PyTorch tools to create smaller videos of 12 frames. In [15], the prediction CNN model does not include the use of OpenFace for feature extraction; instead, it is trained based on micro-expressions it identifies. This FER-2013 dataset contains approximately 35,890 labeled images with 7 emotions: anger, fear, disgust, joy, surprise, sadness, and neutrality.

Specific emotions. If micro-expressions cannot be suppressed when a person lies, then certain emotions could be specific to deception. Facial expressions can be translated into emotions in [15][24][25]. In [25], children's facial expressions are translated into emotions using FACET. Two key moments in the video are analyzed under the assumption that cognitive load can vary, and therefore the expressed emotions are different. The first phase corresponds to the period when the child hears the question asked. They are aware that they will be asked to talk about the event that occurred, specifically about a mistake they did not commit. They will be more curious to learn about what they will be questioned on and will evaluate the situation rather than try to monitor their expressions. During this phase, facial expressions can be particularly informative because children's inhibitory abilities may not yet be activated. The second phase corresponds to the period immediately following the first, once the question is asked in full. The child may be in a decision-making process about what to say to the experimenter and how to formulate their response. This phase is interesting for evaluating the expressions of children who decide to lie by omission.

The emotions of surprise and fear appear in [25] as indicators of deception since they are the most prevalent expressions when comparing emotions that appear in a situation where the child lies versus where they do not lie.

Verbal cues can also be analyzed for deception, including speech rate, frequency of pauses, and inconsistencies [14]. However, facial analysis is more relevant in children because they exhibit fewer verbal cues than adults, which would limit the analyses. In [15][25], it was not the presence of a specific emotion that allowed labeling a video, but rather the sudden variation from one emotion to another and the difference in emotions between different phases. A baseline emotion is identified in [15] for comparison purposes. The dominant micro-expression identified is the neutral expression, which corresponds to moments when the subject is not interacting with the experimenter and is sincere. This expression reflects confidence and comfort when answering questions, with stable emotion levels and no sudden mood variations. In a subsequent phase, the subject was required to both lie and tell the truth, which can generate discomfort and be reflected in emotional variations. The neutral expression is compared to the dominant expression identified during this phase, which is joy in this case. Therefore, the model understands that joy and the shift from neutral to joyful are indicators of deception. The expression of fear is studied in [24], under the assumption that, in high-stakes situations, the subject would fear getting caught. Although fear can be found in an honest person and can therefore be misinterpreted, the degree of suppression will be different between a liar and a truth-teller.

Multimodal Approach. There are several approaches to multimodality, one of which involves studying multiple cues separately, while the other involves considering all cues together as a whole.

FacialCutNet [14] uses a set of indicators: frequency of action units, facial symmetry, and gaze direction. The model consists of several components combining CNN, LSTM, and an attention module. Once the indicators are extracted, FacialCutNet concatenates them and predicts an output value.

Eye Dynamics & Blinking. A key element of facial dynamics is eye movement, i.e., the opening and closing of eyes in each video frame. In [26], eye dynamics, measured by the Eye Aspect Ratio (EAR) (relationship between eye width and height), and the blink frequency, analyzed through histograms of blink percentage, are studied.

Cognitive Load. In addition to analyzing facial expressions, [22] proposes a cognitive analysis by measuring the duration of participants' performance. The assumption is that cognitive load is higher in lying situations as the subject exerts more effort, engaging more significant mental processes than situations where the individual is honest.

Extraction methods and tools. There are several extraction methods and tools that are used in our paper corpus: OpenFace, IDT, CNN, and one proposal creates a custom dataset from existing ones, as shown in table 3.

Table 3. Extraction methods.

Paper	[14]	[15]	[20]	[22]	[23]	[24]	[25]	[26]
OpenFace	x		x	x	x	x	x	x
IDT			x					
CNN		x						
Custom dataset								x

OpenFace. This advanced behavioral analysis tool is designed to extract and analyze facial features from videos. It provides detailed data on the location of facial landmarks, the confidence of the facial recognition algorithm, gaze direction, head posture, as the presence and intensity of AUs, facilitating automated study of facial expressions [29].

IDT (Improve Dense Trajectory). [20] uses handcrafted features extracted with IDT using MBH (motion bountary histogram), HOG (histogram of oriented gradients), and HOF (histogram of optical flow) to extract motion features.

Convolutional Neural Network (CNN). It belongs to deep learning, inspired by the structure of neural networks. They have made significant contributions to computer vision more than any other algorithm and are used in various tasks such as image segmentation, video and image recognition, and image classification.

Custom Dataset. A custom dataset is used in [26] 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). They ran experiments on 3 datasets and generalize them by training on one dataset and transferring it to another.

3.3. Classification of the results

The corpus papers are using several machine learning algorithms, as shown in table 4. Some are famous, like the random forest algorithm or the decision tree one, and some are new proposals, specific for the research field, like the FacialCueNet proposition.

Table 4. Classification methods.

Paper	[14]	[15]	[20]	[22]	[23]	[24]	[25]	[26]
Random Forest					x	x	x	x
Decision Tree					x			
SVM			x		x			
LSTM			x					
CNN		x						x
C3D			x					
Logistic Regression					x			
KNN					x			
Multiple Instance Learning								x
FacialCueNet	x							

Random Forest [30]. This classification algorithm consists of individual prediction trees that form a decision forest. Each tree in the model is a classification tree that predicts an output. The use of Random Forest is motivated by its reliability and stability on small samples in [25], while in [26], it is used for comparison purposes. This model is employed with features obtained by OpenFace as input data to predict deception in [26], whereas in [25], Random Forest is used both to determine which features or combinations of features were relevant and for classification.

Decision Tree [31]. This decision support hierarchical model uses a tree-like model of decisions and their possible consequences. This algorithm is used in [23] and compared to other techniques.

SVM is chosen in [20] because it has shown to be effective for small amounts of data. Two approaches have been used for this classifier: the first relies on extracting facial features from expert-defined features (handcrafted features), while the second uses OpenFace to extract facial features. The obtained data serves as input to SVM for prediction.

Long Short Term Memory (LSTM). LSTMs are a type of neural network that can also classify and process sequential data. In [20], LSTM is used on facial features also extracted by OpenFace. Unlike other models, LSTM can directly process the sequence of AU activations without needing to perform aggregation calculations. The model bases its predictions on all images in the videos and their reciprocal positions. Since training LSTM requires a lot of time, the model was validated only on data obtained during the interrogation phase, as the authors believe it contains more obvious indicators of deception than a free speech phase where the subject is not directly confronted.

CNN. In [15], the model is trained using FER-2013. It compares the differences in emotions present between phases where the subject is truthful and those where they are not.

C3D. The 3D Convolutional Neural Network is designed to process spatiotemporal information by improving the identification of moving images and 3D images. This network automatically extracts features from video clips without the need for prior feature extraction. In [20], although C3D can extract facial features itself, since it receives raw video images as input, OpenFace was used to target only the face in the videos. The dimensions of each image are then normalized to 112 x 112 to feed into C3D. This data preprocessing helps overcome GPU memory constraints.

Logistic regression. This type of statistical model (also known as logit model) is often used for classification and predictive analytics. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds (the probability of success divided by the probability of failure). [23] use this algorithm and compare it to Random Forest, Decision Tree, SVM, and KNN.

KNN. It is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy [32].

Multiple instance learning [33]. MIL deals with problems with incomplete knowledge of labels in training sets. More precisely, in multiple-instance learning, the training set consists of labeled "bags", each of which is a collection of unlabeled instances. A bag is positively labeled if at least one instance in it is positive, and is negatively labeled if all instances in it are negative. The goal is to predict the labels of new, unseen bags. This technique is used in [26].

FacialCueNet. It is a video-based deep-learning network for deception detection that can effectively use instantaneous changes in facial expression and provide spatial and temporal interpretation of prediction results [14].

Technical Constraints. To overcome GPU memory constraints encountered in [20], the video is divided into blocks of 16 images to predict the entire video by aggregating the predictions from previous blocks. This method helps

estimate the average intensity of deception cues throughout the video. The issue of imbalanced classes may also arise. Although the authors in [25] corrected their sample to account for having more liar classes than truth-teller classes during model development, it remains significantly biased towards classifying children as liars rather than truth-tellers.

3.4. Model evaluation

To measure the results of machine learning models, most of the mentioned studies rely on the following metrics: Area Under the Curve (AUC) [14][20], Accuracy [14][15][23][24][25][26], TP rate [24], FP rate [24], Precision [14][15][23][24], Recall [14][15][23][24], F-Measure [24], PRC area [24], Kappa [24], and F1 score [14][15][23]. AUC is equivalent to Accuracy and reflects the model's ability to make correct predictions during the training phase. Accuracy represents the ratio of correct predictions to the total number of predictions. It measures how often a model predicts the correct outcome. Precision represents the number of correctly predicted positive instances by the model. It measures how often a model correctly predicts the positive class. Recall represents the proportion of actual positives that were correctly predicted by the model. F1-Score represents the harmonic mean of precision and recall values. It is a good indicator to evaluate a model's performance. An F1-Score close to 1 indicates that the model is performing well.

Facial expression. For facial expression analysis, the SVM model coupled with OpenFace used in [20][23] seems to demonstrate better performance compared to other prediction models for their respective datasets. Bag of Lies achieves an accuracy of 61.54%. However, other models manage to classify videos correctly with results above 53% for metrics such as accuracy, precision, recall, and F1-score, for most of the models used. For the Holiday dataset, an AUC of 0.72 is achieved for videos without cognitive load, while an AUC of 0.78 is achieved for videos with cognitive load. Overall, classifiers tested in [20] show better performance on videos with cognitive load, suggesting that lying requires more effort than telling the truth, especially when unexpected questions are asked. In an experiment conducted in [20] to evaluate human performance, the results show that gender or education level has no impact on the ability to predict lies, confirming the findings of [25] that gender is not a factor influencing lie detection ability. Furthermore, human performance compared to machine performance decreases, perhaps due to a "fatigue effect" where attention decreases over time. Despite humans relying on multiple criteria such as narrative details, eye contact, and facial expressions, they remain less effective than machines with an AUC of 0.57. On the other hand, [26] is less optimistic about the models' abilities to identify lies. With Random Forest, the results show overfitting on the training data and an inability to generalize on the validation set. Real-life trial showed better performance, with an AUC of 0.58 compared to Crime and Opinion. This could be due to the fact that the features for this dataset were manually annotated, while the features were learned by the model for the other datasets. When the model is trained on Opinion and tested on Crime, its performance is not significantly higher to justify generalizing the model across datasets. This is because prediction models are highly specific to their dataset and cannot be tested on a dataset that varies greatly from the original dataset. Finally, for CNN, its accuracy on these datasets is comparable to a "coin flip."

Emotions. Increased surprise expression among liars can be interpreted as general excitement when prompted to think about transgression (unlike non-liars who haven’t experienced it). It may be more pronounced in liars because it’s the first time they’re being questioned about it. The expression of fear can indicate that the child perceives lying as a moral violation and encounters a conflict between two social obligations: one being to confess to someone that the toys are broken, the other being to remain silent as suggested by someone else. In [15], the prediction model achieves a classification accuracy of 74.17%, which is considered very high compared to other models designed using CNN and the FER-2013 dataset. However, this metric alone does not determine the actual performance of the model. It would be more appropriate to evaluate the F1 score, which is not provided in the study. Additionally, the number of participants does not allow us to conclude that the model is good. In [25], the model predicts lying with consistent accuracy regardless of age, sex, or history of child maltreatment. However, it is observed that their model performs better with children than with adults. The metrics used, such as the average expression score or the expression range to summarize automatic FACS data, can influence the model's ability to effectively detect lies.

Multimodal Approach. The results from [22] showed that deceptive acts lasted longer than truthful acts. Specifically, videos involving the Sophisticated Lie strategy had significantly longer durations than those of the Simple Lie or pure Truth, highlighting an increase in cognitive load associated with the Sophisticated Lie. This difference is particularly pronounced in dynamic non-verbal conditions, where the action time for the Sophisticated Lie exceeded that of the Dynamic Verbal condition. This suggests that media richness influences performance

duration. These findings support the Media Richness Theory, showing that more complex deception strategies require increased cognitive load, reflected in prolonged action durations, especially in contexts where verbal communication is absent. Eye movement and blinking, studied in [26], show that liars tend to exhibit more consistent eye dynamic patterns, while truth-tellers are more relaxed and often show greater eye movement amplitudes. The neural network's performance in detecting lies based on eye dynamics is better on the Opinion dataset, contradicting the cognitive load hypothesis. Regarding eye blinking, predictions from data sequences of the same video clip are aggregated, and the entire clip is predicted as "lie" if the result exceeds a threshold. This threshold is selected to maximize training accuracy with desired false positive rates. Based on the ratio between blink percentages in truth and lie situations, an accuracy of 56.2% is achieved on Crime, which is comparable to human performance. It also show that liars tend to blink less (blink percentage $\leqslant$ 2%) compared to truth-tellers who blink more frequently (blink percentage $\geqslant$ 4%). However, no specific threshold or general probabilistic pattern clearly distinguishes liar groups from control groups.

4. Conclusion

We presented 8 experiments about lie detection from facial expression with machine learning techniques. All studies in the corpus rely on limited data, with the number of participants in the experiments not exceeding 200. Even though the figures show positive precision for prediction models, there is a paradox in which a model may have very high precision but, ultimately, may not be applicable to other contexts because it is too specific to the data on which it was trained. The performance of models in predicting lies is equivalent to or slightly better than human capacity. Precise lie-related features vary from one study to another but can overall help the model learn lying cues and detect them. However, the question of subjectivity and bias still arises. Can we say that the model is reliable when trained on a dataset where the fundamental truth is not always known? Many lie detection studies use real-life trial data, which may have judicial errors and evidence interpretation issues. The results are perhaps not sufficient to justify the use of these ML/DL models in real-world contexts. Moreover, models trained on specific datasets may not generalize effectively to real-life situations where deceptive behaviors vary significantly. In everyday life, individuals may lie in very different ways depending on various factors such as cultural context, personal experiences, and emotional state. For instance, facial micro-expressions, often used to detect lies, can be influenced by contextual factors not present in training data. Therefore, a model that performs well in a controlled environment may fail to detect lies in varied real-world scenarios. To enhance the robustness and reliability of these models, it is crucial to incorporate diverse and representative datasets that reflect the multitude of ways people may lie. Additionally, techniques such as regularization and domain adaptation may be necessary to enable models to better adjust to new conditions encountered in real-world settings.

Note that relying solely on facial expressions to spot lies poses significant challenges, as emotions and their expressions can vary greatly across different cultures. Additionally, individuals may lie in ways that do not manifest in their facial expressions, such as through subtle body language cues or verbal inconsistencies, making it essential to consider a broader range of indicators. To enhance the reliability and effectiveness of lie detection models, several key areas need attention. Firstly, increasing dataset diversity is crucial. This means collecting data from a wide range of cultural backgrounds and real-life scenarios to ensure the model can generalize across different contexts and detect lies regardless of cultural variations in expression. Secondly, enhancing model interpretability is essential. Understanding how the model arrives at its conclusions allows for greater transparency and trust, enabling users to comprehend and validate the decision-making process. Lastly, addressing ethical concerns is imperative. This involves ensuring the respectful and fair use of data, avoiding biases that could lead to unfair treatment or discrimination, and safeguarding the privacy of individuals. By focusing on these areas, we can work towards developing more robust, trustworthy, and ethical lie detection systems.

References

[1] DePaulo, Bella M., James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper (2003) "Cues to deception." *Psychological bulletin* 129(1): 74

[2] DePaulo, Bella M., Deborah A. Kashy, Susan E. Kirkendol, Melissa M. Wyer, and Jennifer A. Epstein (1996) "Lying in everyday life." *Journal of personality and social psychology* 70(5): 979

[3] Bond Jr, Charles F., and Bella M. DePaulo (2006) "Accuracy of deception judgments." *Personality and social psychology Review* 10(3): 214-234

[4] Gongola, Jennifer, Nicholas Scurich, and Jodi A. Quas (2017) "Detecting deception in children: A meta-analysis." *Law and human behavior* 41(1): 44

[5] Lajevardi, Seyed Mehdi, and Zahir M. Hussain (2012) "Automatic facial expression recognition: feature extraction and selection" *Signal, Image and video processing* 6: 159-169.

[6] Porter, Stephen, and Leanne ten Brinke (2010) "The truth about lies: What works in detecting high-stakes deception?" *Legal and criminological Psychology* 15(1): 57-75.

[7] Bhatt, Priya, Amanrose Sethi, Vaibhav Tasgaonkar, Jugal Shroff, Isha Pendharkar, Aditya Desai, Pratyush Sinha et al. (2023) "Machine learning for cognitive behavioral analysis: datasets, methods, paradigms, and research directions." *Brain informatics* 10(1): 18.

[8] Lewis, Jerry A., and Michelle Cuppari (2009) "The polygraph: The truth lies within." *The journal of psychiatry & law* 37(1): 85-92.

[9] Burzo, Mihai, Mohamed Abouelenien, Veronica Perez-Rosas, and Rada Mihalcea (2018) "Multimodal deception detection." *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2*, pp. 419-453

[10] Vrij, Aldert, Katherine Edward, Kim P. Roberts, and Ray Bull (2000) "Detecting deceit via analysis of verbal and nonverbal behavior" *Journal of Nonverbal behavior* 24: 239-263.

[11] Rojas-Barahona, Lina, Bo-Hsiang Tseng, Yinpei Dai, Clare Mansfield, Osman Ramadan, Stefan Ultes, Michael Crawford, and Milica Gasic (2018) "Deep learning for language understanding of mental health concepts derived from cognitive behavioural therapy." arXiv preprint arXiv:1809.00640.

[12] Lu, Haoyu, Qiongyi Zhou, Nanyi Fei, Zhiwu Lu, Mingyu Ding, Jingyuan Wen, Changde Du et al.(2022) "Multimodal foundation models are better simulators of the human brain." arXiv preprint arXiv:2208.08263

[13] Speth, Jeremy, Nathan Vance, Adam Czajka, Kevin W. Bowyer, Diane Wright, and Patrick Flynn (2021) "Deception detection and remote physiological monitoring: A dataset and baseline experimental results" *International Joint Conference on Biometrics (IJCB)*, pp. 1-8

[14] Nam, Borum, Joo Young Kim, Beomjun Bark, Yeongmyeong Kim, Jiyoon Kim, Soon Won So, Hyung Youn Choi, and In Young Kim (2023) "FacialCueNet: unmasking deception-an interpretable model for criminal interrogation using facial expressions" *Applied Intelligence* 53(22)

[15] Yildirim, Suleyman, Meshack Sandra Chimeumanu, and Zeeshan A. Rana (2023) The influence of micro-expressions on deception detection. *Multimedia Tools and Applications* 82(19): 29115-29133.

[16] Jupe, Louise Marie, and Keatley, D. Adam (2020) "Airport artificial intelligence can detect deception: or am i lying?" *Security Journal* 33(4)

[17] Frank, Mark G., and Stennett, Janine (2001) "The forced-choice paradigm and the perception of facial expressions of emotion." *Journal of personality and social psychology* 80(1): 75.

[18] Lu, Guanming, Xiaonan Li, and Haibo Li (2008) "Facial expression recognition for neonatal pain assessment." *International conference on neural networks and signal processing,* pp. 456-460

[19] Ekman, Paul, and Wallace V. Friesen (1978) "Facial action coding system." *Environmental Psychology & Nonverbal Behavior*

[20] Monaro, Merylin, Stéphanie Maldera, Cristina Scarpazza, Giuseppe Sartori, and Nicolò Navarin (2022) "Detecting deception through facial expressions in a dataset of videotaped interviews: A comparison between human judges and machine learning models" *Computers in Human Behavior* 127: 107063.

[21] Kitchenham, B.A. (2004) “Procedures for Undertaking Systematic Reviews”, *Joint Technical Report, Computer Science Department*, Keele University (TR/SE-0401) and National ICT Australia Ltd. (0400011T.1)

[22] Zhou, Xingchen, Rob Jenkins, and Lei Zhu (2023) "An Honest Joker reveals stereotypical beliefs about the face of deception" *Scientific reports* 13(1): 16649.

[23] Islam, Siam, Popin Saha, Touhidul Chowdhury, Asif Sorowar, and Raqeebir Rab (2021) "Non-invasive deception detection in videos using machine learning techniques." *International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, 1-6.

[24] Shen, Xunbing, Gaojie Fan, Caoyuan Niu, and Zhencai Chen (2021) "Catching a liar through facial expression of fear" *Frontiers in Psychology* 12: 675097.

[25] Bruer, Kaila C., Sarah Zanette, Xiao Pan Ding, Thomas D. Lyon, and Kang Lee (2020) "Identifying liars through automatic decoding of children's facial expressions" *Child development* 91(4): e995-e1011.

[26] Belavadi, Vibha, Yan Zhou, Jonathan Z. Bakdash, Murat Kantarcioglu, Daniel C. Krawczyk, Linda Nguyen, Jelena Rakic, and Bhavani Thuriasingham (2020) "MultiModal deception detection: Accuracy, applicability and generalizability." *TPS-ISA conference*, pp. 99-106

[27] Thomson, S. B (2011) “Qualitative Research: Validity”. *JOAAG*, 6(1)

[28] Sartori, Giuseppe, Sara Agosta, Cristina Zogmaister, Santo Davide Ferrara, and Umberto Castiello (2008) "How to accurately detect autobiographical events." *Psychological science* 19(8): 772-780.

[29] Baltrusaitis, Tadas, A. Zadeh, Y. C. Lim, and L. P. Morency (2018) "Openface 2.0: Facial behavior analysis toolkit” *IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*" pp 59-66

[30] Ho, Tin Kam (1995) “Random Decision Forests”. *International Conference on Document Analysis and Recognition*, Montreal, pp. 278–282

[31] von Winterfeldt, Detlof, and Edwards, Ward (1986) "Decision trees". *Decision Analysis and Behavioral Research*. Cambridge University Press. pp. 63–89. ISBN 0-521-27304-8.

[32] Hastie, Trevor (2001) “The elements of statistical learning : data mining, inference, and prediction”, Springer

[33] Babenko, Boris (2008) "Multiple instance learning: algorithms and applications"