Development of an Enhanced Machine Learning Model for Deception Detection Leveraging Facial Action Units and Linguistic Features

Adetoye Oluwatoyin Adedokun ADEDOKUNADETOYE@GMAIL.COM
Department of Computer Science
University of Ibadan,
Ibadan, Oyo State, Nigeria.

Adebola K. Ojo ADEBOLAK.OJO@GMAIL.COM
Department of Data Science
University of Ibadan,
Ibadan, Oyo State, Nigeria.

Editor: Sakinat Folorunso, Roseline Ogundokun, and Francisca Oladipo

Abstract

High-stakes deception in human interaction has become a prevalent social activity today. These actions trigger changes in the deceiver's behavior and physiological responses, potentially offering clearer signs of dishonesty. However, traditional deception detection techniques often rely on subjective human judgment and may not consistently produce reliable results. Recently developed Deception Detection models also trained their models using simulated datasets. The task of deception detection however remains challenging because no single indicator can reliably predict deception on its own. There is currently no empirical evidence demonstrating that any specific pattern of physiological responses is uniquely indicative of deception.

This study proposes a bimodal machine learning approach for multiclass deception detection by integrating facial Action Unit (AU) features and linguistic features extracted from a real-life dataset of video interviews of suspected criminal perpetrators, persons of interest in an ongoing investigation, and convicted criminals. Facial expression features were obtained using the Facial Action Coding System (FACS), while linguistic cues were derived from the speech transcripts of participants using the Linguistic Inquiry Word Count (LIWC) scores. The extracted features were combined using a mid-level data fusion strategy to form a unified feature representation. A Random Forest classifier was used to multi-classify communication instances into three categories: Truth, Lie, and Uncertain. Our dataset, consists of 3720 samples collected from real-life interactions, was divided using an 80:20 train–test split, and 10-fold cross-validation was applied during the training process to improve model reliability. In addition, hyperparameter tuning using Grid Search was performed to determine the optimal configuration of the classifier.

The results of our experiment shows that our proposed bimodal deception detection model achieved an overall classification accuracy of 88%. Further evaluation using metrics such as precision, recall, F1-score, confusion matrix, and multiclass ROC curves confirms the effectiveness of our proposed approach.

Our findings show that the combination of facial and linguistic cues from real-life dataset provides a richer representation of deceptive behavior and improves classification performance. This study therefore contributes to the development of automated deception detection systems based on mid-level fusion and machine learning multiclassification of real-life dataset.

Keywords: Bimodal Machine Learning, Multiclass Classification, Mid-level Data Fusion, Random Forest.

1. Introduction

Deception is a common social behavior in which individuals intentionally mislead others, conceal the truth, or promote false beliefs for personal gain. It can occur in different forms, ranging from low-stake deception, where the consequences are minor (such as exaggerated product reviews or harmless pranks), to high-stake deception, where the outcomes can be severe or life-altering, including wrongful convictions, job loss, financial fraud, or even loss of life. As a result of the its impact on the society, deception detection has become an area of interest and an important research area in various fields like criminal investigations, human resources and recruitment, business transactions, and panels of inquiry. Traditional deception detection techniques, including polygraph tests, Functional Magnetic Resonance Imaging (FMRI), and electroencephalograms (EEG), have previously been widely studied. However, their approaches have several limitations: they often require the use of specialized equipment, trained professionals, and controlled laboratory environments, making them costly, time-consuming, and difficult to deploy widely. Furthermore, some of these techniques can raise ethical, privacy, and consent issues because they involve measuring, capturing or scanning a person's physiological or brain signals and activities. Due to the technological advancement in Machine Learning, there has been an introduction of data-driven approaches that can identify patterns and making predictions from large and complex datasets. Machine learning models can be leveraged to automate complex tasks and adapt them to different conditions, making them a viable tool for deception detection. There have been previous machine learning studies that relied on generic machine learning algorithms designed for binary classification and these algorithms have been used to train simulated datasets, which may limit the generalization ability and realism of the results.
As a solution to these limitations, we propose a machine learning approach for detecting deception in physical interviews by integrating facial and linguistic cues extracted from a real-life dataset. The dataset consists of video interviews of suspected criminal perpetrators, persons of interest in an ongoing investigation, and convicted criminals. The datasets were collected from publicly available interviews on YouTube. For ethical reasons, we ensured that the data is used strictly for academic purposes and does not harm or misrepresent the individuals involved. We also ensured there were no disclosure of unnecessary personal information of the interviewees and there were no identifiable details during analysis and reporting to protect privacy.
The proposed model performs multi-class classification that classifies responses into Truth, Lie, and Uncertain. The findings of this research contribute to improving automated deception detection and have potential applications in areas such as criminal investigations, border security screening, and investigative panels where accurate assessment of truthfulness is essential.

2. Related works

Tsuchiya et al. (2023) proposed a machine learning–based method for identifying deceptive responses in remote interviews by correlating facial expression features with variations in the interviewees' pulse rate. The drawbacks of this research were that they used few/limited dataset and they could not capture enough different cultural background and neuro-divergent statuses. Also, pulse as a deception detection modality can be manipulated of controlled by the interviewee. An experimental assessment of the proposed method was conducted using a 10-fold cross-validation strategy with a Random Forest classifier. The findings indicate that, across individual subjects, both accuracy and F1-score generally ranged from 0.75 to 0.80, with peak performance reaching 0.87 for accuracy and 0.88 for the F1-score.
Recent advances in automated deception detection have increasingly focused on the integration of verbal and non-verbal cues to improve classification performance Dinges et al. (2024). Early studies emphasized the complementary role of linguistic and behavioral features in identifying deceptive behavior Abed et al. (2020), while subsequent research explored the effectiveness of advanced machine learning algorithms such as Random Forest variants and CatBoost for classification tasks Aldania et al. (2023). Ensemble-based approaches have also demonstrated promising results by leveraging facial cues and combining multiple classifiers to enhance detection accuracy Avola et al. (2021). More recently, bimodal frameworks integrating diverse feature extraction techniques have been proposed to capture both verbal and visual indicators of deception Bahaa et al. (2024). Furthermore, end-to-end deep learning architectures have shown the potential to automatically learn high-level representations from video data, reducing dependence on handcrafted features Dinges et al. (2023). Collectively, these studies highlight a growing trend toward multimodal and ensemble-based deception detection systems aimed at improving robustness, accuracy, and generalizability across different application domains. Yildirim et al. (2023) trained a deep learning model to analyze facial changes and extract meaningful information using the Facial Expression Recognition 2013 dataset. Based solely on nonverbal cues, they developed a deception detection classifier that achieved an accuracy of 74.17%. They noted that incorporating additional modalities, such as vocal features and textual data alongside video features, would be necessary to create a fully autonomous system in the future.

3. Proposed method

We propose a Bimodal machine learning framework for deception detection that integrates facial cues and linguistic patterns extracted from real life interview recordings. This method combines facial features derived using Facial Action Coding System (FACS), and Linguistic Inquiry and Word Count (LIWC) tool. These bimodal features were fused using a mid-level data fusion strategy and subsequently used to train a machine learning classifier for deception detection. The general framework of our proposed system is shown in Figure **??**fig:figure1-deception-detection-system-architecture].

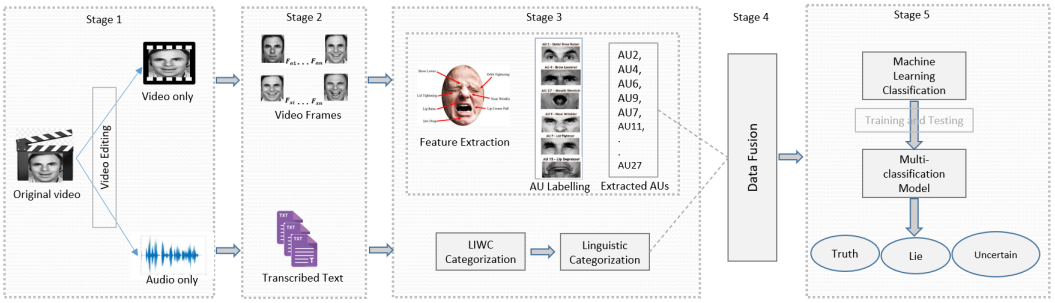

Figure 1: Deception Detection System Architecture

3.1. Data Collection

The proposed method was developed to identify deception in high stake scenarios, through physical interviews captured in videos using the machine learning approach. High stakes scenarios in the context of this research refers to any event in which a successful deception by a perpetrator, or a false detection of deception by the interlocutor may have dire consequences such as jail time, death sentence, loss of job or other serious consequences. To obtain realistic behavioural cues, audio-visual datasets were collected from publicly available real-life interview recordings on YouTube. The dataset consisted of 89 video recordings obtained from 89 participants. These videos include interviews involving criminal suspects, witnesses, victims' relatives, or individuals connected to criminal incidents. Such interviews often occur during police interrogations, media interviews, or investigative journalism reports. The use of real-life datasets was motivated by the need to capture genuine behavioural and linguistic patterns associated with deception. Many earlier studies such as Tsuchiya et al. (2023) relied on simulated datasets, where volunteers were instructed to fabricate stories or deliberately exhibit deceptive behaviour. Our dataset exhibited diversity in demographic qualities by including participants of different age groups, genders, and ethnic backgrounds, which helped improve the robustness and representativeness of the dataset for machine learning analysis. Each participant corresponded to one video recording, resulting in a total of 89 videos, with an average duration of approximately 5 minutes per video. Simulated environments on the other hand often fail to reproduce the psychological pressure and emotional states present in real investigative situations. By utilizing real-life interviews, this study aims to improve the validity of the dataset and enhance the model's capability to detect deception in realistic contexts. The decision to use a real-life dataset was motivated

by the need for high-quality data capable of improving the training and overall accuracy of the model compared to most previous researches that used simulated data where volunteers were asked to act-out a lying scenario or deliberately exhibit deceptive behaviours.

3.2. Data Preprocessing

Our videos datasets were originally recorded by media organizations, law enforcement agencies, or forensic professionals. These recordings were not originally produced for machine learning analysis, so they contained irrelevant segments such as introductions, advertisements, or unrelated discussions. To address this issue, the videos were manually edited to remove unnecessary sections to ensure that only the segments relevant interview were retained for analysis. We then separated the videos recordings into audio and video streams to allow each modality to be processed independently. The extracted video streams were decomposed into individual frames. To increase the number of training samples, each interview video was segmented into multiple response units corresponding to individual statements made by the participant during the interview. Each response segment was annotated and labeled as Truth, Lie, or Uncertain. The corresponding annotations were done by verifying each statement with the ground truth and evidence already established by the law enforcement and forensic professionals. The segmentation process expanded our dataset to 3,720 labeled instances used for machine learning training and evaluation. These frames were analyzed to detect facial movements and expressions. Using facial analysis techniques based on the Facial Action Coding System (FACS), relevant Action Units (AUs) corresponding to facial muscle movements were extracted and labelled into a category of 16 AUs (AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU15, AU17, AU20, AU23, AU25, AU26, AU45). The audio stream was transcribed into text using speech-to-text techniques. The resulting transcripts were subsequently analyzed using Linguistic Inquiry and Word Count (LIWC) to generate linguistic and psychological features. The categories of the linguistic features are Negative Emotion, Anxiety, Anger, First person pronoun, Positive Emotion, Psychological Processes, Personal Concerns, Cognitive process.

3.3. Feature Extraction

Feature extraction was performed separately for each modality in order to capture relevant behavioural and linguistic indicators of deception.

3.3.1. Facial Feature Extraction (FACS)

Facial behavioural cues were extracted using the Facial Action Coding System (FACS). FACS is a broadly used framework for analyzing facial expressions by tracking the activation of individual facial muscles, referred to as Action Units (AUs). AUs are the smallest movements of the face which can be reliably distinguished from another. Each AU is associated with one or more muscle movements. Individual video frames were analyzed to detect the presence and intensity of specific action units associated with facial movements per frame. These action units were converted into numerical feature values representing the activation intensity of facial muscles.

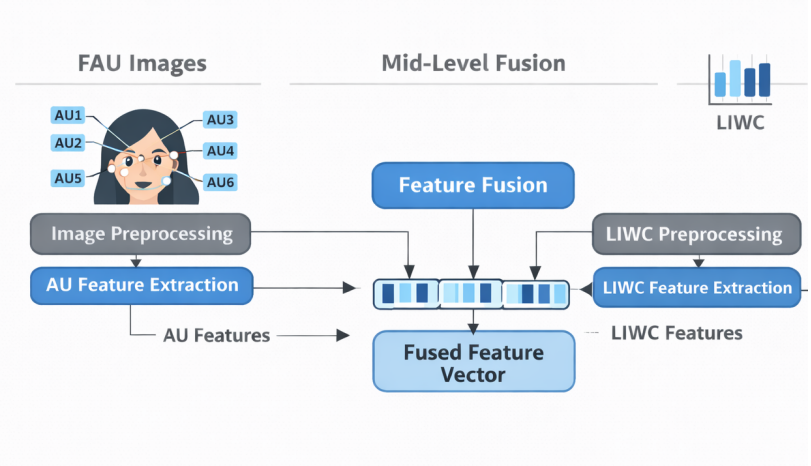

Figure 2: Mid-Level data fusion

3.3.2. Linguistic Feature Extraction (LIWC)

The transcribed textual data were processed using the Linguistic Inquiry and Word Count (LIWC) framework to extract the linguistic features. The textual transcript was first normalized (punctuation handling, lowercase conversion), after which it was tokenized and then mapped to LIWC dictionaries. The frequency of words in each category was computed including: affective processes (positive/negative emotion, anxiety, anger), cognitive processes (insight, causation, discrepancy, certainty), pronoun use (first-person singular/plural, third-person), and function words (articles, prepositions). The features were aligned with the AUs for a bimodal fusion.

3.4. Feature Normalization

Because the extracted features originated from different modalities with varying numeric ranges, normalization was required to ensure uniformity. Min–Max normalization was applied to scale all features into a common range.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

This transformation ensures that features that have larger numeric values do not dominate the learning process. Feature normalization also improves model convergence and training stability.

3.5. Data Fusion

We carried out Mid-Level data fusion to merge the attributes which are from two different data sources into a unified representation, which is subsequently used to train machine learning models. The fusion process is described in Figure 2. The concatenated features form a combined feature vector, which was then used as input for the machine learning classifier.

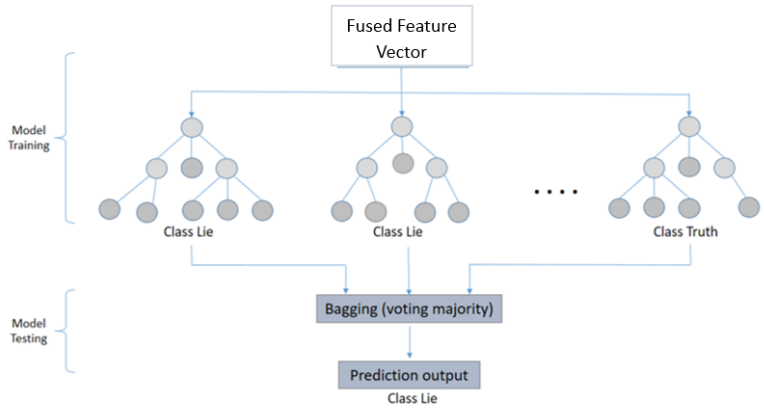

Figure 3: Random Forest Multiclassification

3.6. Machine Learning Model Training

The bimodal feature vector was used to train the Random Forest classifier. Random Forest classifier was selected due to its robust ensemble learning technique, ability to handle high-dimensional datasets, and resistance to overfitting. The classifier was trained to categorize interview responses into three classes - Truth, Lie, and Uncertain, representing different deception states. Our dataset consisted of 3,720 labeled instances, with the following class distribution: Truth (1,500 instances), Lie (1,440 instances), and Uncertain (780 instances). This class distribution shows the relative occurrence of truthful, deceptive and uncertain responses within the collected dataset. We used the fused bimodal feature vectors corresponding to these labeled instances to train the Random Forest model shown in Figure 3 and learn the associated patterns of each deception class.

3.7. Performance Evaluation

After our model has been trained, we further analyzed the classification performance of the proposed deception detection model, using performance metrics such as precision, recall, accuracy and confusion matrix. The calculations for these metrics are defined below.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

$$F1Score = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{5}$$

Figure 4: Running analysis for deception detection

4. Experimental Results

This section presents the results of our experiment obtained from the implementation of the proposed machine learning bimodal deception detection model described in the methodology. The objective of the experiment was to evaluate the effectiveness of combining facial and linguistic indicators in detecting deceptive behaviour during real-life interview scenarios.

We implemented the deception detection model using the Python programming language within the Anaconda development environment. We utilized several machine learning and data processing libraries, including scikit-learn, NumPy, and other supporting libraries.

4.1. Model Training and Testing

After a successful training of the deception detection model, we got an overall classification accuracy of 0.88. In addition to the accuracy gotten, we computed other performance evaluation metrics like precision, recall, and F1-score to provide a comprehensive assessment of the performance of the model as shown in Table 1. We evaluated the trained model using a separate test dataset. Once the testing script is executed, the system launches a Graphical User Interface (GUI) that allows users to perform deception analysis on video data as shown in Figure 4 As shown in Figure 4, our model performs an analysis on the loaded video and the prediction is made as the video is played; the confidence per prediction is also calculated and displayed and the prediction history is generated with their corresponding time stamps and confidence level.

4.2. Prediction Log and Confidence Score

After the analysis of a video is completed, the system generates a prediction log file containing detailed results of the analysis. This file is stored in CSV format and provides a chronological record of predictions for the entire video. Each log entry contains: timestamp, predicted class, transcribed speech, Confidence score, feature values used for prediction.
Example log entry:

4:22:30, lie, "I was at home all day", 0.723

4.3. Experimental Results and Discussion

The results of the performance of the proposed deception detection model are presented in Table 1. Our model was evaluated using a 10-fold cross-validation. We evaluated the model performance using metrics such as accuracy, precision, recall and F1 Score as shown in Table 1.

Table 1: Deception Detection Model Performance

	Precision	Recall	F1 Score	Support
Uncertain	0.77	0.89	0.87	1,500
Lie	0.84	0.83	0.8	1,440
Truth	0.85	0.73	0.77	780
Accuracy			0.88	3,720
Macro Average	0.82	0.82	0.81	3,720
Weighted Average	0.90	0.88	0.86	3,720

The model achieved an overall accuracy of 0.88. This shows a strong classification performance for detecting deceptive behavior.
The Receiver Operating Characteristic (ROC) analysis as shown in Figure 5 was performed using the One-vs-Rest approach. The ROC curve shows a visual tradeoff between the True Positive Rate (TPR) and False Positive Rate (FPR) across different thresholds of the classification. The Area Under the ROC Curve (AUC) value was 0.90 for class Truth, 0.89 for class Lie and 0.87 for class Uncertain. A summary of the comparative analysis of our model with other previous studies using metrics such as classifier used, modalities, type and dataset, number of subjects, accuracy and key characteristics is shown in Table 2.

5. Conclusion

Real-life datasets have shown to be more effective in deception detection as opposed to simulated datasets which does not accurately train the model to detect deception. Also, our machine learning multiclassification creates a new class of "Uncertain" different from class "Lie" which reduces the chances of wrong classification into the "Lie" class and helps our model to classify more accurately.
In the course of this research, we discovered that deception may not be absolute, that is, a conversation may not entirely be deceptive. We may therefore not be able to generalize an entire conversation as either truthful or deceptive, instead, we can break the conversation into segments of deception and truthfulness.
This research has proven that machine learning can be successfully utilized for deception detection. We recommend that this work be applied in real life scenarios such as criminal investigation, border security, panels of enquiry and in any other field where there is a need for High-stakes deception detection.

Table 2: Model Comparative Analysis

Study	**Method**	**Classifier**	**Modalities**	**Dataset**	**Number of Subjects**	**Accuracy**	**Key Characteristics**
Tsuchiya et al. (2023)	Machine Learning Classifier	Random Forest	Facial Expressions + Physiological Signals (Pulse Rate)	Experimental deception dataset	32	83%	Combines facial expressions with physiological measurements to improve deception detection accuracy
Yildirim et al. (2023)	Micro-Expression Analysis using Machine Learning	Convolutional Neural Network	Facial Micro-Expressions	Controlled facial micro-expression dataset	40	79%	Focuses on the role of subtle facial micro-expressions in identifying deceptive behavior
Dinges et al. (2024)	Artificial Intelligence-based Deep Learning Model	Convolutional Neural Network	Facial Cues (Video)	Facial deception video dataset	52	86%	Uses automated facial cue extraction and AI models to detect deception in video recordings
This Study	Random Forest with Mid-Level Fusion	Random Forest	Facial Action Units + Linguistic Features	Real-Life Bimodal Dataset	89	88%	Bimodal fusion of real-life dataset improves detection of truth, lies, and Uncertain

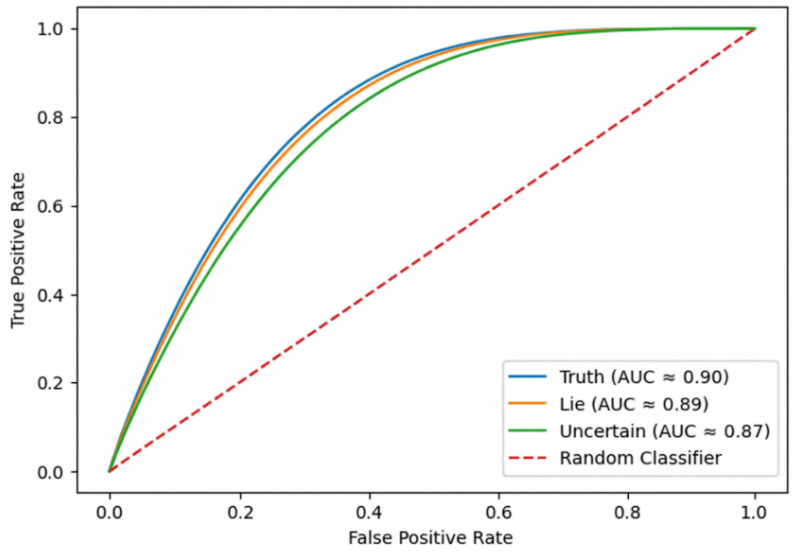

Figure 5: Multiclass ROC Curve (One-vs-Rest) Analysis

Future research could expand on this work by examining how cultural, racial, and gender differences affect action units (AUs) and the accuracy of deception detection.

References

Shaimaa H Abed, Ivan A Hashim, and Ali Sadeq A Jalal. Verbal and non-verbal features in deception detection systems. In *2020 3rd International Conference on Engineering Technology and its Applications (IICETA)*, pages 78–83. IEEE, 2020.

Annisarahmi Nur Aini Aldania, Agus Mohamad Soleh, Khairil Anwar Notodiputro, et al. A comparative study of catboost and double random forest for multi-class classification. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 7(1):129–137, 2023.

Danilo Avola, Marco Cascio, Luigi Cinque, Alessio Fagioli, and Gian Luca Foresti. Lietome: An ensemble approach for deception detection from facial cues. *International Journal of Neural Systems*, 31(02):2050068, 2021.

Mohamed Bahaa, Mena Hany, and Ehab E Zakaria. Advancing automated deception detection: A multimodal approach to feature extraction and analysis. In *International Conference on Intelligent Systems, Blockchain, and Communication Technologies*, pages 727–738. Springer, 2024.

Laslo Dinges, Marc-André Fiedler, Ayoub Al-Hamadi, Thorsten Hempel, Ahmed Abdelrahman, Joachim Weimann, and Dmitri Bershadskyy. Automated deception detection from videos: using end-to-end learning based high-level features and classification approaches. *arXiv preprint arXiv:2307.06625*, 2023.

Laslo Dinges, Marc-André Fiedler, Ayoub Al-Hamadi, Thorsten Hempel, Ahmed Abdelrahman, Joachim Weimann, Dmitri Bershadskyy, and Johann Steiner. Exploring facial cues: automated deception detection using artificial intelligence. *Neural Computing and Applications*, 36(24):14857–14883, 2024.

Kento Tsuchiya, Ryo Hatano, and Hiroyuki Nishiyama. Detecting deception using machine learning with facial expressions and pulse rate. *Artificial Life and Robotics*, 28(3):509–519, 2023.

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