scientific reports

Check for updates

OPEN

Multimodal machine learning for deception detection using behavioral and physiological data

Gargi Joshi[1], Vaibhav Tasgaonkar[1], Aditya Deshpande[1], Aditya Desai[1], Bhavya Shah[1], Akshay Kushawaha[1], Aadith Sukumar[1], Kermi Kotecha[1], Saumit Kunder[1], Yoginii Waykole[1], Harsh Maheshwari[1], Abhijit Das[2], Shubhashi Gupta[2], Akanksha Subudhi[2], Priyanka Jain[2], N. K. Jain[2], Rahee Walambe[1,3]✉ & Ketan Kotecha[1,3]✉

Deception detection is crucial in domains like national security, privacy, judiciary, and courtroom trials. Differentiating truth from lies is inherently challenging due to many complex, diversified behavioural, physiological and cognitive aspects. Traditional lie detector tests (polygraphs) have been widely used but remain controversial due to scientific, ethical, and practical concerns. With advancements in machine learning, deception detection can be automated. However, existing secondary datasets are limited—they are small, unimodal, and predominantly based on non-Indian populations. To address these gaps, we present *CogniModal-D*, a primary real-world multimodal dataset for deception detection, specifically targeting the Indian population. It spans seven modalities—electroencephalography (EEG), electrocardiography (ECG), electrooculography (EOG), eye-gaze, galvanic skin response (GSR), audio, and video—collected from over 100 subjects. The data was gathered through tasks focused on social relationships and controlled mock crime interrogations. Our multimodal AI-based score-level fusion approach integrates diverse verbal and nonverbal cues, significantly improving deception detection accuracy compared to unimodal methods. Performance improvements of up to 15% were observed in mock crime and best friend scenarios with multimodal fusion. Notably, behavioural modalities (audio, video, gaze, GSR) proved more robust than neurophysiological ones (EEG, ECG, EOG).The study demonstrates that multimodal features offer superior discriminatory power in deception detection. These insights highlight the pivotal role of integrating multiple modalities to develop robust, scalable, and advanced deception detection systems in the future.

Keywords Multimodal data fusion, Automated deception detection, Lie detection, Affective computing, Cognitive behaviour analysis, Neurophysiological data, Behavioural data

Human behaviour and cognition are inherently multimodal and occur through the integration of multiple heterogeneous modalities of verbal factors such as pupil dilation, eye blink, eye tracking, brain activity, thermal imaging, micro expression and nonverbal cues, facial expressions[1], audio, video[2] and text[3] presenting a comprehensive and concise picture of the world around us[4]. Deception detection is an alarming and lasting social challenge that potentially impacts human lives. Deception detection is deciding whether a particular communication carries the truth through thoughts, feelings, and intentions[5]. The task is even more challenging and complex as no single clue alone is an indisputable predictor of deception[6]. The accurate detection of deceptive behaviour with malicious intent is far more crucial for several mission-critical applications such as law and interrogating agencies, courtroom trails, and national security purposes covering physiological (e.g., biosensors and thermal imaging), visual (e.g., facial expressions and gestures), speech (e.g., pitch and pause length), and linguistic modalities[7]. Deceit/ lie can be an acted lie, deliberately lying under duress, and lying through partial truths[8]. Lack of details, change in vocal pitch tone, fillers, contradictions, inconsistencies, reduced speech, eye contact, pupil size, and rigid body are behaviour-based definitive signs of deception that must be traced accurately[9]. Existing work in the domain applies techniques such as polygraph, eye tracking, speech, emotion, visual, acoustic, physiological, linguistic, and social dynamics, exploring cognitive aspects of deception[10].

[1]Symbiosis Institute of Technology, Symbiosis International Deemed University, Pune, India. [2]Centre for Development of Advanced Computing (C-DAC), Delhi, India. [3]Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International Deemed University, Pune, India. ✉email: rahee.walambe@sitpune.edu.in; rahee.walambe@gmail.com; director@sitpune.edu.in

Consequently, a range of recent work focuses on employing data-driven approaches based on machine learning and deep learning algorithms and techniques[11] that can accurately detect deception based on several multimodal cues from audio, video, and textual modalities rather than relying on a single modality for detection[12]. The interplay between verbal and nonverbal cues for the data captured from human subject stimuli through various sensors joint modelling and integration of neuro modalities improve the underlying task performance of deception detection compared to the reliance on individual modality as a whole[13]. Despite a lot of progress in the field of automatic deception detection datasets and methods, there persist critical problems that need further investigation in terms of addressing language and cultural gaps, scarcity of real-life labelled data with open access availability, larger datasets and benchmarks to build robust and generalizable deception detection systems addressing the cultural, ethical and security concerns altogether[14]. Although multiple secondary datasets[15–19] exist for multimodal deception detection, they do not explore the cognitive and behavioural aspects of deception. These datasets are primarily collected from subjects with Caucasian ethnicity. Such datasets do not generalize well for other ethnic populations; hence, a dataset is needed to focus specifically on the Indian population. In[15], the authors present a Bag-of-Lies dataset of video, audio and eye-gaze data collected from 35 Indian participants. In this work, the aim was to identify the individual and combined contribution of various non-verbal (video, gaze), behavioural (audio) and neurophysiological modalities (EEG, ECG, EOG, GSR) cues collected from a large Indian population. To that end, this manuscript addresses the potential gap of the non-availability of a very rich multimodal large dataset focusing on the Indian population. The experimental study is executed on a one-of-a-kind novel, comprehensive, real-world multimodal primary dataset, consisting of data from over 100 subjects and spanning seven diverse verbal and nonverbal modalities like EEG, ECG, EOG, Gaze, GSR, Audio, and Video. This work demonstrates an improved task performance for generalized accurate and robust multimodal deception detection.

In summary, the key contributions of this research are:

1. Development of a comprehensive multimodal dataset called CogniModal-D for deception detection collected from over 100 Indian participants balanced across age and gender. The work presents an experimental study and detailed discussion on the design of the experimental design paradigm for primary data collection involving various tasks and procedures followed for collecting (EEG, ECG, EOG, GSR, gaze, audio, and video) targeting the Indian population with a sample size of over 100 subjects, along with the annotation process.
2. The study, comparison and analysis of various unimodal ML models applied to the CogniModal-D dataset to explore their efficacy and other performance parameters, deepening the understanding of more complex fine-grained concepts through feature extraction and classification techniques on various deception detection tasks and scenarios.
3. Implementation and demonstration of the effectiveness of the multimodal machine learning-based fusion approach for the individual deception detection tasks in comparison to their unimodal counterparts, highlighting the overall efficacy of the investigated modalities and the adopted strategy. The contribution and importance of the verbal, nonverbal, neurophysiological, and cognitive aspects of multimodal deception detection are explored.

The paper is organized as follows: Sect. 2 discusses the previous work and related studies on machine learning-based unimodal and multimodal deception detection. Section 3 presents the methodology adopted and employed for experimental design for primary data collection, followed by Sect. 4, where automated deception detection with multiple modalities is presented. The results are presented, analyzed, and discussed in Sect. 5. The conclusion and future scope of the work are presented in Sect. 6.

Related work

The literature survey was carried out mainly in three aspects: deception detection paradigms, Machine Learning –Deep Learning-based techniques developed explicitly for deception detection, and Multimodal fusion approaches for cognitive behaviour analysis and deception detection. Traditional machine learning algorithms such as support vector machines, random forests, k-nearest neighbours, and decision trees are used for deception detection[5]. Deep learning-based models such as Convolutional Neural Networks, Recurrent Neural Networks and Long Short Term Memory, Multilayer Perceptron, Multi-scale CNN, Deep Belief Networks, and Multimodal Auto Encoders have been widely applied by previous studies[20]. Multimodal data fusion processes combine diverse modalities to achieve accurate and robust predictions. Depending on the level at which the modalities are fused in the model architecture, the fusion approaches applied are broadly classified into early, late, and hybrid fusion[21].

i) Feature-level fusion or early fusion: Features from different modalities are combined into a single feature vector, capturing correlations and associations among modalities[17,22–24].
ii) Decision-level fusion or late fusion: Individual modality-specific architectures are built, trained, and combined to have an outcome. The score-level approach is applied in the late fusion and generates a score separately for each modality, indicating the likelihood that it belongs to a particular class. Late fusion is the widely applied fusion technique[15,25,26]. The methods for fusing multimodal features include major concatenation[17], score-level fusion[15,25], and average and majority voting mechanisms[27].
iii) Intermediate or hybrid fusion: Hybrid fusion provides a more flexible approach to combining modalities by integrating the benefits of both early and late fusion[3,18,28,29].

Deception detection tasks span across diverse domains such as psychology, linguistics, computer science, and neuroscience[16]. Consequently, a range of recent work focuses on the task of multimodal deception detection. In[17], the verbal and nonverbal visual, acoustic, and linguistic cues are explored. Deception detection based on emotional states and facial features is explored in various works[30,31]. In[32], multimodal deception detection through conversational dialogues with facial and linguistic behaviour is reported. Deception detection in videos with audio, text, and nonverbal features is presented by Zhang et al.[33]. The power of combined features of PPG, eye tracker, audio, and video for mock crime scenarios is exploited[19]. High-level features for deception detection with real-life trial data with acoustic, visual, and textual modalities are explored[22]. The gender-based variability of multimodal deception detection in males and females is studied in[34]. Multimodal deception detection with audio, visual, and textual cues is performed in[20]. The significance of of speech and nonverbal cues in lying are explored in[35]. The combination of verbal with physiological cues aids in improved deception detection as reported in[36]. The importance of facial cues and emotions is studied for lie detection in[37,38,23]. Deep learning-based deception detection through visual features is carried out in[39]. The comparison of low-stake and high stake lies with machine learning is carried out in[40]. Deception detection from gaze and speech using a multimodal attention LSTM-based framework is performed in[41]. Voting-based method for automatic deception detection from videos using audio, visual, and linguistic features is proposed in[27]. Multimodal features for detecting political deception are explored in[18]. Linguistic and physiological data streams are fused with bimodal CNN[42]. A dataset for conversational dialogues with multimodal features with cross modal features is studied in[43]. The automated deception detection from verbal, visual and videos is presented in[44]. Recently, audio-visual deception detection has been studied in[24,26]. A study on non-invasive approaches to multimodal deception detection based on multimodal features such as facial expression, body movement, audio, video, and thermal imaging is presented in[45]. EEG based mental state detection is carried in[46]. The nonverbal cues for deception, such as silence and speech, are explored in[47]. Multimodal deception datasets are analyzed and evaluated in[21]. Although a lot of studies are carried out on this topic, they are not in the context of the Indian population, and yet there is a need for a real-world comprehensive multimodal dataset that undermines, exploits, and analyzes the verbal, nonverbal, physiological, and cognitive aspects of multimodal deception detection.

CogniModal-D: experimental design protocol and primary data collection

Dataset description

The CogniModal-D dataset developed as part of this is a comprehensive multimodal dataset built for efficient and robust deception detection, exploring cognitive aspects of deception through well-designed tasks. The dataset consists of multiple diverse modalities such as EEG, ECG, EOG, gaze, GSR, audio, and video, exploiting the verbal and nonverbal cues of deception. Over 100 participants' samples belonging to the Indian population from various strata/backgrounds were collected and tested as a part of the experimental study. The data is realistic and age and gender-neutral to address the inherent bias in the data collection process. The 100 sample set is approximately equally divided into male and female participants. The participants are healthy controls recruited and selected through open advertisement, fulfilling the inclusion criteria defined. The participants are categorized into three age groups, namely- 18–25, 25–40, and 40–60 years respectively. Each age group has an equal number of male and female participants. All methods were carried out in accordance with relevant guidelines and regulations as per the ethical approval obtained from Symbiosis International Deemed University's (SIU) Institutional Ethics Committee with approval number SIU/IEC/517. The study and procedure along with its use for research purpose was explained to all participants and informed consent was obtained from all subjects.

The sample distribution details are listed in Table 1 below.

Experimental tasks for data collection

The deception detection paradigm aims at identifying and detecting deceptive behavior. Deceit detection through multiple modalities induces behavioral and physiological changes in the behavioural patterns which can be captured effectively by various sensing mechanisms. For example, pupil dilation and erratic eye gaze (Eye tracking sensor and EOG), increased sweating (GSR), change in voice and facial expressions (Audio and video data), increased heart rate (ECG), change in the brain waves (EEG) which can be very effectively leveraged for

Modality	Audio, Video, Gaze, GSR		EEG, ECG, EOG
Records	100		100
Subjects	100		100
Truth-Lie	50-truth 50-lie		50- truth 50-lie
Age Groups	18-25	25-40	40-60
Number of Participants	33	34	33
Gender Groups	Female – 53		Male - 47

Table 1. Dataset sample distribution details.

accurate deception detection. In this study, we designed two specific tasks- the best friend scenario and the mock crime scenario- which the participants must undertake. The data for the mentioned seven modalities was collected while the participants were carrying out these tasks in controlled lab settings. One of the challenging tasks in the primary dataset development is to develop the experiment tests so that the participants can generate the data corresponding to a particular class. In the specific application of deception detection, generating and collecting data corresponding to lie (class 0) and truth (class 1) is essential. During data collection, appropriate annotations must also be provided. In this work, we created two scenarios of mock crime physical task derived from[7] and a best friend scenario task[15,16].

Mock crime

The mock crime task consists of controlled mock crime interrogations. Existing mock crime tasks deal with virtual crimes[7]. In contrast, in this work, a physically realistic scenario is presented to the participant. The participant completes the task as instructed and is asked to describe their performance. Randomly, 50% of the total participants (group 1) are asked to carry out specific misbehavior such as stealing, harming the property, etc. The remaining 50% (group 2) are instructed to carry out the task normally and honestly. The participants are not monitored and are isolated during this task. After the task completion, the group 1 participants are asked a few leading questions which they need to answer in such a way as to fool the system to obtain high scores or incentives by lying or deceiving the system by hiding information or some evidence, lying about stealing some assets or trying to fool/confuse the system. In the mock crime paradigm, the participant admits lies/deceives, or fools the system to voluntarily achieve high scores or incentives. Group 2 is asked to answer truthfully. Group 1 samples are annotated as class 0 or negative class and group 2 as class 1 or positive class. While they carry out this task, the data from the aforementioned seven modalities is collected and labelled accordingly.

Best friend scenario

The best friend task consists of storytelling about social relationships[16]. The best friend scene impacts the emotional complexity and relational dynamics influencing deceptive behavior. The participant is requested to think of their best friend/closest relative and talk about good things (positive truth) and bad things (negative lie) about them. Later, they are asked to remember a person they dislike and describe that person positively (positive lie) and negatively (negative truth). Data is available for these four classes and then scaled down to two classes, lie and truth, by merging positive truth and negative truth into the truth class and positive lie and negative lie into the lie class.

Significance of modalities and their acquisition

The verbal, nonverbal, and neurophysiological modalities such as EEG, ECG, EOG, GSR, gaze, audio, and video are captured simultaneously through various sensor devices attached to the participants.

Electroencephalogram (EEG) Data Modality: EEG is a highly valuable modality in neuroscience, providing a direct measure of electrical activity in the brain with exceptional temporal resolution. It captures fluctuations in this activity through electrodes placed on the scalp, translating them into waveforms representing the brain's neural dynamics. Its non-invasive nature allows for safe and repeated measurements, crucial for long-term studies and monitoring. One can significantly gain insights into the cognitive processes underlying deceptive practices by analyzing specific EEG markers, such as frequency band alterations and time-frequency-based features. EEG signals can be combined with other bio-signals such as ECG, EOG, GSR, and Eye tracking signals for multimodal fusion methods for detecting lies, anomalies, and suspicious behaviour. The data was acquired using a 28-channel EEG device. The data format was stored across three types of files: a brain vision header file (.vhdr), the EEG data file (.eeg), and a marker file (.vmrk). Figure 1 depicts the EEG cap and system montage. The spatial distribution of the EEG electrodes follows the 10–10 system to ensure consistent and comparable EEG recordings across different participants.

Electrooculography (EOG) Data Modality: EOG is a non-invasive technique that measures the electrical activity generated by the eye with the help of electrodes. The eyes produce electrical potentials due to the movement of the eyeball. These signals can be captured and analyzed to gain insights into various cognitive and physiological processes. In cognitive behavioural analysis, EOG is often employed to study attention, cognitive processes, and emotional responses by analyzing eye movements and their underlying patterns. By tracking eye movements, information about individuals right from where individuals focus their attention during tasks provides insights into cognitive processes such as reading, problem-solving, and decision-making. EOG is also used to investigate how emotions impact visual attention and exploration. Changes in eye movements and gaze patterns are associated with deceptive behavior. While EOG provides valuable information, it is often used with other physiological measures and behavioural assessments to understand cognitive and emotional processes comprehensively. Interpreting EOG data requires expertise in both the technical aspects of signal processing and the psychological or cognitive context under investigation.

The data was acquired using a 2-channel EOG setup. The data format was stored across three types of files: a brain vision header file (.vhdr), the EOG data file (extracted from .eeg), and a marker file (.vmrk). Figure 2 shows the EOG channels used for data collection, where four electrodes are shown. For this work, we collected data from EOG_{left} and EOG_{right} locations only, as the eye tracking modality is also collected separately.

Electrocardiography (ECG) data modality: ECG is a diagnostic modality that measures and records the heart's electrical activity. This technique involves placing electrodes on the patient's skin in specific locations to capture electrical signals produced by the cardiac muscle during each heartbeat. The recorded data are presented as a waveform on an ECG machine, which healthcare professionals analyze to identify normal and abnormal heart rhythms, diagnose heart conditions, and monitor cardiac health. For deception detection, ECG is used as part of polygraph tests; ECG measures similar physiological responses, theorizing that lying induces stress-

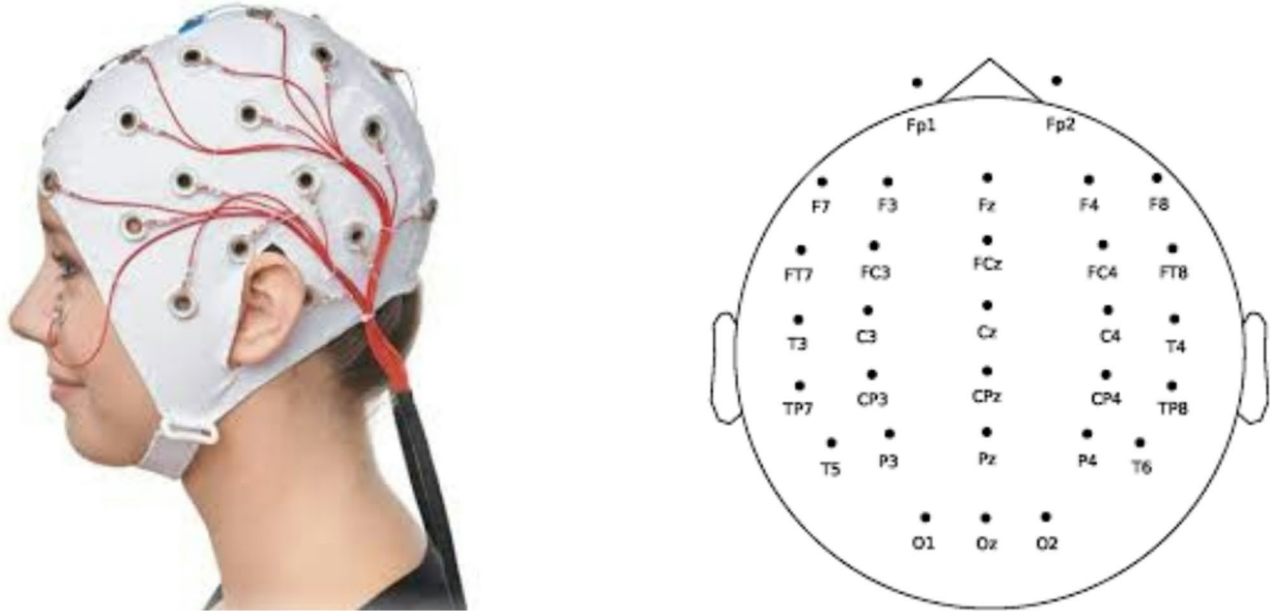

Fig. 1. (**a**): EEG cap. (**b**): EEG electrode placement.

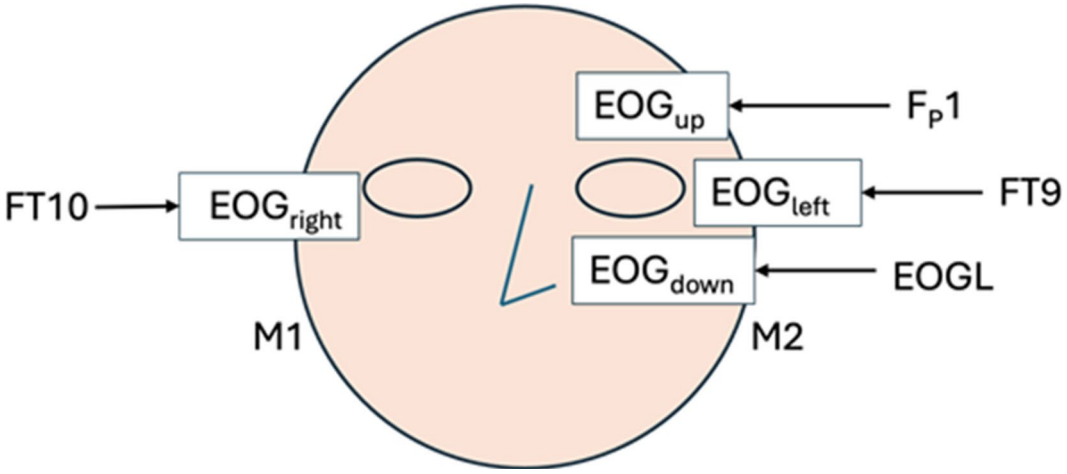

Fig. 2. Connection points for EOG channels for data collection.

related changes in cardiac activity. The data was acquired using a 2-channel ECG setup. The typical electrode placements for ECG were positioned to capture cardiac activity effectively. These electrodes were placed to mimic standard ECG lead placements, although exact positions were specified by the manufacturer. The first derivation of the Einthoven Triangle was used. The Bipolar Lead Configuration is employed where the first ECG electrode is placed on the right chest area, close to the heart (e.g., below the right clavicle) (RA), and the second is placed near the left chest area (LA) position. The ground electrode was available in the EEG cap. This common reference electrode acted as a virtual ground and helped reconstruct the third missing lead using differential calculations following the Einthoven Triangle principle.

The data format was stored across three types of files: a brain vision header file (.vhdr), the ECG data file (extracted from the .eeg file), and a marker file (.vmrk). Figure 3 shows the ECG electrodes and the channel connections for data collection.

Audio modality: Audio plays a crucial role in deception detection as changes and alterations in the voice pitch modulation and tone, verbal cues in the form of words and language, nonverbal cues such as the use of fillers and inconsistencies, physiological and emotional manipulations, changes in the vocal characteristics, patterns verbal and nonverbal patterns in speech are potential indicators of deceptive practices that are traced and detected through audio analysis. A common approach to building a training model for audio is to use a deep neural network (DNN) that incorporates various layers, such as convolutional layers, dense layers, and activation layers. The model learns features at different time scales by applying filters with small receptive fields to the audio

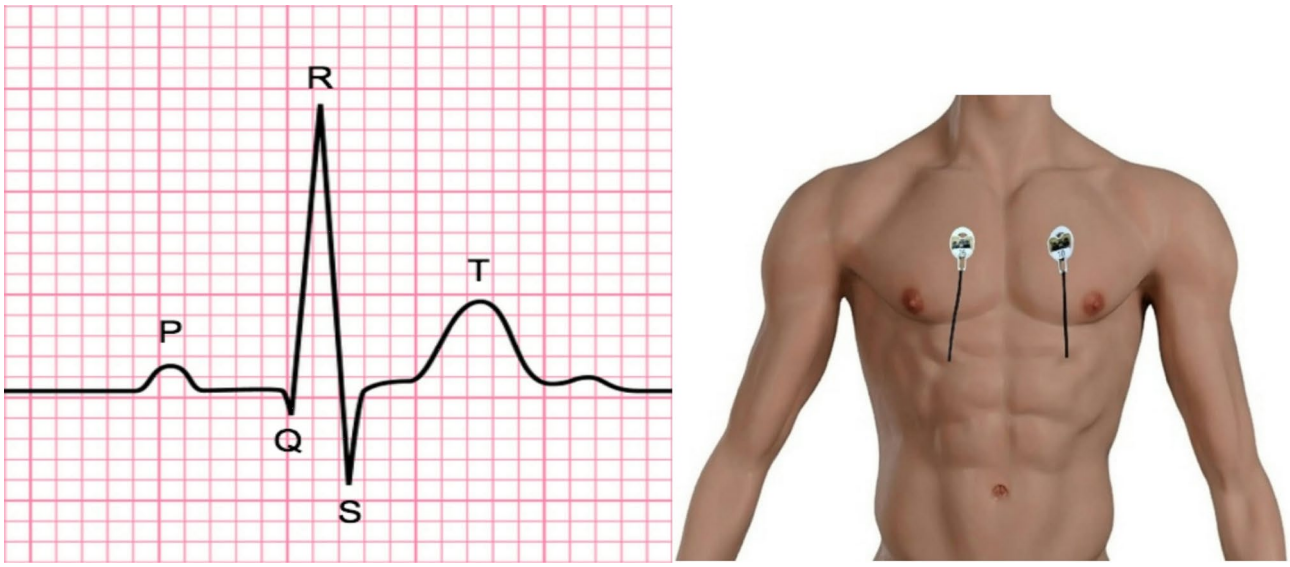

Fig. 3. PQRST complex and placement of the ECG electrodes.

data. The output layer of the sequential DNN model typically consists of one or more neurons, depending on the specific audio classification task. The choice of activation function in the output layer depends on the task at hand. The sequential DNN learns to recognize and classify various audio patterns by adjusting the model's parameters during training. The final layer of the sequential DNN model is the output layer, which typically consists of one or more neurons, depending on the specific audio classification task. The sequential DNN learns to recognize and classify various audio patterns by iteratively adjusting the model's parameters during training.

Video data modality: Video plays a significant role in deception detection by decoding body language and facial expressions in the form of micro-expressions, gestures, and eye contact. Contextual dependencies are well captured through the video data. Physiological manipulations and emotional appeals, physiological and behavioral analysis with nonverbal cues such as facial expressions and inconsistencies in body language. Thus, video plays a multifaceted role in leveraging visual and auditory elements for identifying potential deceptive behavioral traces.

Gaze or eye tracking data modality: Eye gaze provides nonverbal cues for deception in the form of eye movement, blink rate, erratic and fast unconcentrated graze, etc. Increased blink rates are indicators of nervousness and increased cognitive load. Aversion in gaze indicates deception; pupil dilation, micro-expressions, and contextual dependencies are considered in conjunction with a broader set of verbal and nonverbal cues. Advanced technologies provide comprehensive behavioral analysis and multiple cues for deception.

Galvanic skin response (GSR) data modality: GSR is a physiological measurement that assesses the electrical conductance of a person's skin. It is primarily influenced by changes in sweat gland activity, which is affected by emotional arousal, stress, or excitement. GSR is utilized to detect potential deception. When individuals experience heightened emotional arousal or stress, their sweat gland activity increases, leading to decreased skin resistance and a rise in GSR readings. In the context of deception detection, GSR has been incorporated into polygraph or lie detection tests, as it is assumed that individuals exhibit increased stress or anxiety when they lie, causing changes in GSR. However, it's crucial to note that GSR alone is not a foolproof method for identifying deception, and its interpretation should be complemented by other physiological measures and behavioural observations to enhance accuracy. Figure 4 shows the placement of the GSR device on hand.

Data acquisition and processing

The data pertaining to EEG, ECG, EOG, and audio and video modalities is collected from the brain vision recorder software. The Gaze and GSR modalities are captured for each subject through Tobi Pro lab software. Raw data is uploaded to the analyzer software, and brain vision analyzer software derives integrated CSV data for all the modalities captured. The integrated CSV file is provided as input to the Python script, which generates markers for each task for synchronisation among the modalities. Each modality has a different sampling frequency e.g. EEG, ECG and EOG are captured at 250HZ, GSR Sampling frequency rate is 60 Hz and eye gaze and eye tracking Sampling frequency rate: 60 Hz.Hence, frequency matching is carried out prior to any pre-processing. A file with integrated data with markers is input for the LSTM model. Data is mounted on Google Colaboratory. Data is available in four classes: Positive truth, negative truth, positive lie, and negative lie. Four classes are further scaled down to two classes, lie and truth, by merging positive truth and negative truth into the truth class and positive lie and negative lie into the lie class. Data is appended into a dictionary of the respective class. The dictionaries are further merged into nested lists. Sequences are generated from the nested list. Labels for respective classes are generated using a 'for loop' over the dictionaries. The individual modalities are processed through their modality-specific architectures, such as Convolutional Neural Networks and Long Short-Term Memory. Diverse data-driven techniques are applied to analyze the performance of each modality.

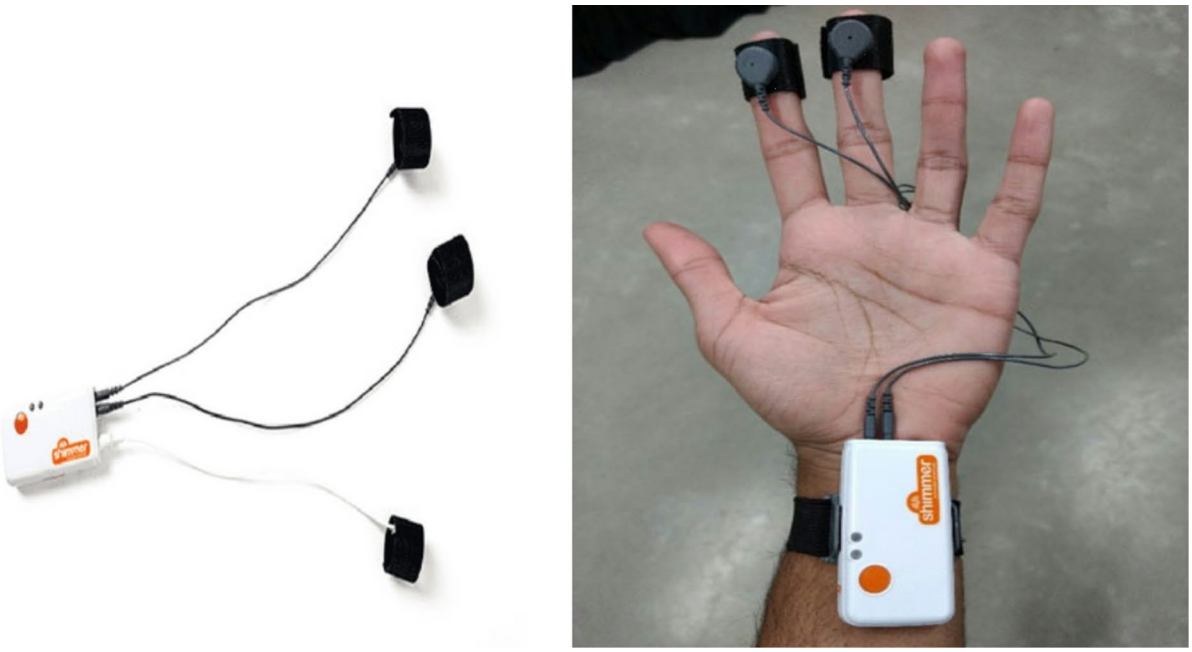

Fig. 4. Placement of the GSR device.

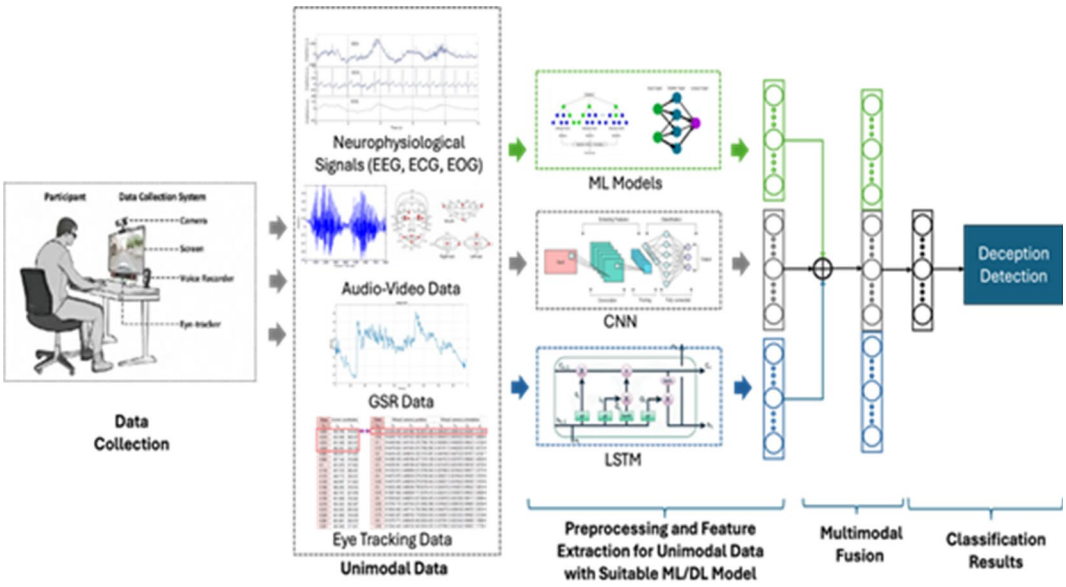

Fig. 5. Multimodal deception detection pipeline.

Unimodal and multimodal deception detection

The multimodal deception detection pipeline is shown in Fig. 5.

Data preprocessing and feature extraction

Unimodal EEG Data preprocessing involves identifying and interpolating bad channels, re-referencing, signal filtering, and artefact removal. The EEG data is initially processed to identify any bad channels that could distort the analysis. The identified bad channels were then interpolated to correct form for any irregularities in the data. The EEG data is re-referenced to standardize the measurement baseline across all channels. A band-pass filter ranging from 1 to 48 Hz was applied to the signal to isolate the frequency bands of interest while excluding noise. In the frontal region of the brain, EEG signals are most commonly associated with the beta frequency band, which typically ranges from 13 to 30 Hz; this means the dominant frequency in the frontal area is usually within the beta range, signifying active cognitive states like focused attention or mental exertion. During relax and stress states, there is significant variations are observed in alpha and beta bands in the frontal regions. While beta

is dominant, other frequencies; namely, theta (4–8 Hz) and alpha (8–12 Hz) can also be observed in the frontal region depending on the mental state. To allow the consideration of all the frequency ranges associated with alpha, beta and theta frequency bands, a bandpass filter ranging from 1 to 48 Hz was applied. The Independent Component Analysis (ICA) algorithm was employed to remove non-brain artefacts from the EEG data, such as those caused by eye movements or muscle activity.

The Feature Extraction process involved analysing key frequency bands (delta, theta, alpha, beta, and gamma). These are firstly extracted, particularly from the frontal regions, due to their significant variations in alpha and beta bands during relaxation and stress events. The Power Spectral Density (PSD) values for these bands were calculated using the Short-Time Fourier Transform (STFT) algorithm. Additionally, the alpha and beta power distribution differences were validated through a repeated ANOVA test to ensure reliability. Statistical entropy was extracted alongside Hjorth parameters, including mobility and complexity, providing insights into the signal's statistical nature and temporal evolution. Statistical features such as mean, median, standard deviation, and kurtosis are computed to capture the central tendency and variability within the EEG data. The EEG modality processing pipeline is depicted below in Fig. 6.

The EOG Data pre-processing consisted of applying A band-pass filter ranging from 1 to 48 Hz to the individual channel (EOG_{Left}, EOG_{right}) signals to isolate the frequency bands of interest while excluding noise. A 50 Hz Notch filter is applied to remove power-line noise. Vertical and horizontal EOG signals are extracted to interpret various eye movements. A high pass filter with a cutoff frequency of 0.1 using a Butterworth filter was used to remove baseline drift. Also, a fourth-order Butterworth band pass filter is applied in the 0.1–20 Hz range. EOG blink peaks, and EOG mean rate (blink) are the main features of eye blinking and are used for further analysis. Statistical features such as mean, standard deviation, and kurtosis are captured to capture the central tendency and variability within the data.

A 0.5 Hz high-pass filter to remove low-frequency trends caused by patient movements or breathing is used for the ECG data. A notch filter at 50–60 Hz (depending on the geographical location) is suggested by the manufacturer to remove electrical interference by significantly attenuating or removing noise frequencies, specifically at 50 Hz or 60 H. A low-pass filter is used to remove high-frequency noise while allowing lower frequencies to pass through. The fluctuations in heart rate occurring at low frequencies (typically between 0.04 and 0.15 Hz) are often associated with sympathetic nervous system activity.

The PQRST wave (Refer Fig. 3) is a component of an ECG reading. It provides information about the heart's function. In many ways, the PQRST wave is the heart's signature, which provides insight into its function and health. Apart from R-peaks, detecting the entire QRS complex can provide more detailed features for stress analysis. The heart rate from the R-R intervals can be computed and can provide a direct indication of stress. The data was segmented into one-minute intervals, and time domain features, including mean RR interval, standard deviation of RR intervals (SDRR), root mean square of successive differences (RMSSD), etc., are extracted. The frequency domain features such as Power spectral density, ultra-low frequencies (0.0 to 0.0033 Hz), very low frequencies (by default, 0.0033 to 0.04 Hz), low frequencies (by default, 0.04 to 0.15 Hz), high frequencies (by default, 0.15 to 0.4 Hz), very high frequencies (by default, 0.4 to 0.5 Hz), total spectral power, ratio of low frequency power by the high frequency, log transformed high frequency etc. can also be extracted. Although not employed in this study, the Non-linear Features such as Poincaré plot parameters, entropy measures, etc., can be used to capture the non-linear nature of heart rate variability.

Audio from all the videos is extracted and processed further to calculate various frequency-based properties such as zero crossing rate, spectral centroid, spectral bandwidth, spectral roll-off, chroma frequencies, and MEL frequency cepstral coefficients (MFCC). The DNN is optimized with a learning rate of 0.0001, trained for 1000 epochs, with a train: test: validation split of 80:10:10 and a batch size of 32.

For the behavioural data, namely audio and video, video recordings of one hundred eleven persons were recorded, with each person contributing four videos (total of 440 videos) about the best friend scenario. To facilitate further analysis, the video files are converted into audio format. Subsequently, Mel-Frequency Cepstral Coefficients (MFCC)[48] was employed as a feature extraction technique to capture relevant information from the audio signals. The extracted MFCC features were organized into a structured dataset, precisely a data frame,

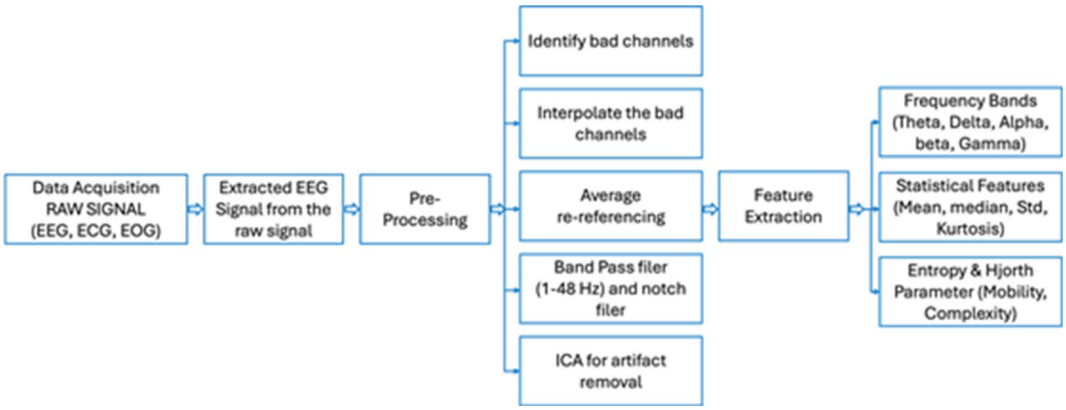

Fig. 6. EEG processing pipeline.

to facilitate efficient handling and manipulation of the data. MFCC, or Mel Frequency Cepstral Coefficients, is a widely used representation for analyzing audio signals and performing tasks such as speech recognition. MFCCs are derived from the mel-frequency spectrum of an audio signal and capture essential characteristics of the spectral shape and envelope of the signal. From the original video data from 110 individuals, 20 frames per individual are extracted. Each individual contributed four videos portraying truth and lies in various scenarios. After frame extraction and preprocessing, a Convolutional Neural Network model was subsequently trained on the data. A multi-step methodology was adopted to comprehensively examine the truth and lies within a video dataset across various conditions. Following previous studies, preprocessing techniques, such as Local Binary Pattern (LBP), enhance the video data's quality and informative content. LBP helps capture local patterns and texture information, making it a valuable tool for feature extraction. Subsequently, a Convolutional Neural Network (CNN) architecture is applied to analyze the preprocessed dataset. For learning hierarchical features and patterns, CNN was employed to discern and classify the underlying visual cues and patterns that distinguish between truth and lies in the videos. Splitting the dataset into 80% training and 20% testing sets ensures a robust evaluation framework. CNN is employed to extract and classify the subtle visual cues and patterns that differentiate truthful and deceptive behaviours. This strategy not only facilitated the identification of these behaviours but also contributed to the robustness and reliability of the model.

Gaze is a crucial modality that plays a vital role in deception detection[15]. Lie and truth data are separated into two folders. To avoid disturbing the data sequence and process each CSV as a single data point. Data preprocessing is performed for a binary classification task. CSV files are read from two folders, irrelevant columns are dropped, and the data and labels are stored in lists. The data points are padded or truncated to have the same number of rows, and the data and labels are converted to Numpy arrays. The long short-term memory (LSTM) architecture is used with a data split of 80 – 20.

For the GSR signal, the preprocessing step involves first subtracting the baseline level from the GSR signal to eliminate any DC offset, effectively normalizing the data to represent the skin conductance in a relaxed state. Although not explicitly specified, removing power-line interference (e.g., 50–60 Hz) is crucial to ensure the accuracy of the GSR signal, especially in regions where power grids produce much noise. Median smoothing and other smoothing techniques are then applied to reduce high-frequency noise and rapid fluctuations in the signal, making it easier to identify and analyze skin conductance responses (SCRs). One of the most crucial steps is decomposing into phasic and tonic components. Filtering at 0.05 Hz effectively removes low-frequency noise and baseline drift from the GSR signal, allowing us to focus on the more rapid changes associated with SCRs. The raw EDA signal is passed through a median value smoothing filter, which helps remove areas of rapid change. The phasic component of the signal is calculated by subtracting the smoothed signal from the original. The convex optimization (cvxEDA) approach aims to separate the tonic (slow-changing) and phasic (rapid-changing) components of the EDA signal more effectively by modelling the underlying physiological processes. Sparse Non-Negative Deconvolution (SparsEDA) is a sparse deconvolution technique that separates EDA's tonic and phasic components. It's based on the idea that SCRs are sparse and can be estimated by deconvolving the observed signal with a predefined basis function.

The training data is fed into a sequential model with two LSTMs, two dropouts, and one dense layer. LSTM model is fitted with L1 and L2 regularizes or lasso and ridge regularizes. The model is trained over ten epochs with 'early stopping' as callbacks, monitoring validation loss. The learning rate applied is 0.0001 with ten epochs, a validation split of 0.1, and a batch size of 32.

Performance evaluation of unimodal machine learning

Various metrics are used to evaluate the performance of multimodal deception detection, ranging from the classification accuracy to the area under the precision-recall curve (AUC). The F1 score, which combines the precision and recall scores, represents the area under the precision-recall curve (AUC) over the test set. Precision refers to the fraction of positive results among the obtained results. At the same time, recall, also known as sensitivity, is the fraction of positive results that were retrieved and were also used as a model performance assessment and evaluation measure. The mock crime and best friend paradigm results are listed below in Tables 2 and 3, respectively.

Multimodal fusion

Multimodal fusion integrates or fuses different heterogeneous modalities to improve the multimodal deception task performance. The individual models are combined through multimodal AI-based fusion techniques such as late or decision-level fusion with CNN-LSTM-based architectures. The modalities are gradually fused, starting with unimodal, to analyze and ascertain the contribution and impact of individual modalities and the combined impact of the fusion of all the modalities in integration on the deception detection task. The modalities are combined through a score-level late fusion approach.

It can be represented mathematically as:

$$Class = argmax\ (S_f, C)$$

Where S_f, C is the combined score of class C

$$S_f = \sum_{i=1}^{n} W_i S_i$$

Where i represents a number of samples, i.e. 100.

n represents the number of modalities, i.e. 7.

Modality	Model	Testing accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
EEG	XG Boost	53	52.7	53	52.84
EOG	SVM	55.5	56	55.5	55.74
ECG	Random forest	69.4	70.2	64.3	67.12
Audio	DNN	43	43	99	59.95
Video	CNN	**96**	**98**	**93**	**94.93**
Gaze	LSTM	64	78	64	70.30
GSR	LSTM	76	56	75	64.12
EEG + EOG + ECG + GSR	Score level fusion of best-performing algorithms on various modalities late fusion	59.3	59.6	57.6	59.75
Audio + Video		50	25	50	33.33
Gaze + GSR		63.4	52	65	57.7
Audio + Video + Gaze + GSR		**75**	**56**	**75**	**64.12**

Table 2. Classification results for unimodal and multimodal data for the mock crime paradigm. Only the Best model results are presented. Significant values are in bold.

Modality	Model	Testing accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
EEG	SVM	48.7	48.7	48.7	48.7
EOG	KNN	53.8	53.8	53.8	53.8
ECG	Light GBM	63.6	66	65	63.15
Audio	DNN	55	60	55	57.39
Video	CNN	**74.1**	**63.9**	**43.3**	**51.6**
Gaze	LSTM	63	71	94	80.89
GSR	LSTM	58	56	94	70.18
EEG + EOG + ECG + GSR	Score level fusion of best-performing algorithms on various modalities late fusion	55.3	56.1	56.1	55.1
Audio + Video		54.4	54.2	88.9	67.4
Gaze + GSR		63	58	90	70.54
Audio + Video + Gaze + GSR		**79**	**65**	**63**	**63.98**

Table 3. Classification results for unimodal and multimodal data for the best friend paradigm. Only the Best model results are presented. Significant values are in bold.

W_i is the weight is assigned to the i^{th} sample, and S_i is the score assigned to the i^{th} samples.

The prediction score for the i^{th} sample belonging to class C as given by modality n. The contribution and impact of modalities are analyzed individually and in fusion.

The multimodal fusion results on the mock crime and best friend paradigm are listed in Tables 2 and 3, respectively.

Results and discussion

The integration of diverse verbal, nonverbal, and neuro-physiological modalities, such as EEG, EOG, ECG, GSR, audio, video, and gaze within a multimodal framework, holds profound significance in the domain of cognitive-behavioural analysis. This holistic approach enables a comprehensive understanding of human behaviour, particularly deception detection. By leveraging the unique insights provided by each modality, a nuanced perspective on the intricate interplay between physiological responses and cognitive processes is established. The comparative analysis for individual modality and multimodal fusion results on the mock crime and best friend task is depicted below in Fig. 7.

From Tables 2 and 3 it can be seen that the EEG modality has an accuracy of 53% with the XGBoost classifier for the mock crime and 48.7% with a support vector machine classifier for the best friend paradigm. EEG unveils neural patterns associated with stress that contribute to lying. It is observed that EOG modality is not capable of detecting the right class as the extracted features are not clearly indicative of the patterns with only ~ 50%+ accuracy for both scenarios. The ECG provides insights into cardiovascular correlates. The ECG modality has an accuracy of 69.4% with a random forest classifier for the mock crime and 63.6% with the Light GBM model for the best friend paradigm, showing a potential role in deception detection. The comparison between the three neurophysiological modalities indicates that ECG is the most indicative of the three signals. This correlates well with the fact that when a person lies, their heart rate increases. Audio recordings offer a rich dataset for behavioural analysis. The audio has an accuracy of 43% for the mock crime paradigm and 55% for the best friend scenario

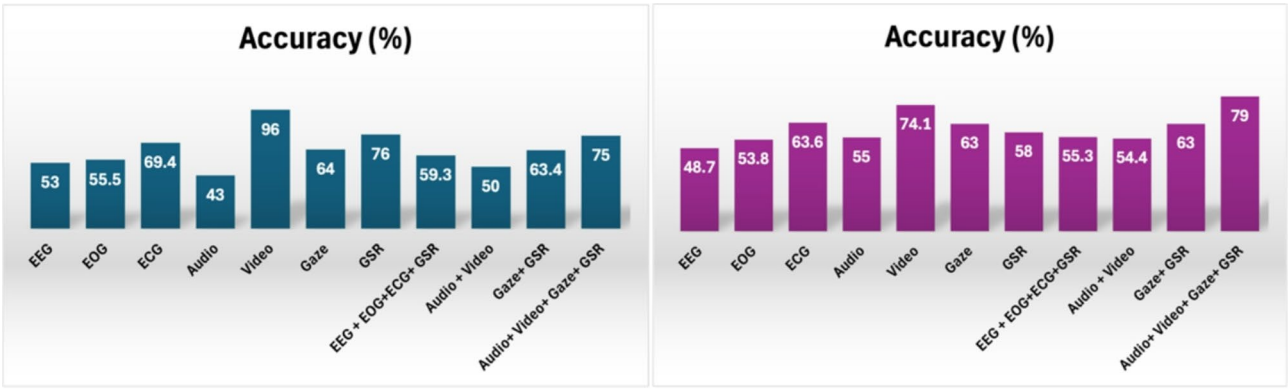

Fig. 7. Results for Individual modality and multimodal fusion for the mock crime task (left) and best friend task (right).

with a sequential DNN. Although audio signals are strong indicators and have multiple features contributing to the behavioural pattern, the models are not able to identify between the varying pitch and other extracted features for both classes. The model demonstrates only about ~ 50% accuracy, demonstrating that the it is not able to classify between two classes. The video modality offers an accuracy of 96% for the mock crime being the highest contributing modality and 74.1% for the best friend paradigm with a convolutional neural network. The facial expressions are very clear indicators of the behaviour. When the person lies, the facial expressions change and clearly indicate the cognitive behaviour. GSR captures changes in skin conductance with an accuracy of 76% for mock crime and 58% for the best friend paradigm with an LSTM model. This correlates with sweating when the person lies. Eye tracking further adds a layer of precision by mapping visual attention. The gaze modality provides an accuracy of 64% for the mock crime and 63% for the best friend paradigm, respectively, with the LSTM model. The video modality is the highest contributor towards mock crime, whereas the gaze modality is vital for the best friend paradigm, followed by the neuro and GSR modalities. The combined score level fusion of EEG + ECG + EOG + GSR has an accuracy of 59.3% for the mock crime and 55.3% for the best friend paradigm. These are neurophysiological parameters only and indicate that the combined effect does not show significant improvement over the contributing unimodal modalities. The fusion of audio and video achieves an accuracy of 50% for mock crime and 54.4% for the best friend paradigm. Although the individual accuracy of video modality is high, the combined accuracy is reduced. The fusion of GSR + gaze provides an accuracy of 63.4% for mock crime and 63% for best friend, signifying that a complementary pairing of modalities is beneficial for the task and is more significant than their counterparts, establishing that multimodal data fusion results in improved performance and better accuracy. The combined multimodal fusion of audio, video, GSR, and gaze provides an accuracy of 75% for the mock crime and 79% for the best friend paradigm, which is improved in comparison to their contributing unimodal data, making the fusion model more robust and accurate at the underlying task of deception detection. It is important to note that the accuracy of unimodal video data for mock crime tasks is 96%. However, it is reduced to 75% when fused with the other behavioural modalities. This is because the contradictory features of the individual modalities bring the accuracy down. However, this makes the fusion model more robust and reliable.

Discussion

It has emerged that behavioural patterns are more valuable indicators and outperforms the multimodal fusion results of the neurophysiological modalities. For example, the Audio + Video + GSR + Gaze combined model outperforms the EEG + ECG + EOG + GSR combined modalities by approximately 15%. The data collection for all the behavioural modalities is non-invasive, less time-consuming, and easier to collect than the neurophysiological modalities, which need longer duration for the EEG cap connections and are much more prone to noises and external artefacts. Traditionally, EEG has always been the best indicator of deception. However, in our experiments, it is observed that the non-neurophysiological modalities, which focus on behavioral patterns, demonstrate better outcomes. The behavioural modalities directly capture readily observable external expressions of emotion like facial expressions, voice tone, and eye gaze, which are easier to detect and interpret than the subtle brainwave patterns measured by EEG/ECG/EOG. The physiological modalities can be easily influenced by external factors and are more subject to individual variability; making audio and video data more reliable and easier to analyze for cognitive behaviour analysis.

When bimodal, tri-modalities, and multiple modalities are fused simultaneously, the results improve further and are better than using a modality individually or in pairs of modalities. Furthermore, using all the modalities, the results are better than using any subset of the modalities. Thus, multiple modalities are highly beneficial for the task, emphasizing the fact that verbal, nonverbal, and physiological modalities combined contribute to and boost the overall performance of the deception detection task by reducing false positives and false negatives, making the model more robust at detection.

The integrated approach not only enhances the sensitivity and specificity of behavioral assessments but also fosters a more comprehensive understanding of cognitive processes in real-world scenarios in providing

distinguished deceptive characteristics, ultimately contributing to advancements in the field of psychology, neuroscience, and human-computer interaction. The choice of better classifiers, architectures, and features plays a paramount role in multimodal deception detection.

Conclusion and future scope

Multimodal deception detection is a crucial task that can be solved by integrating and fusing multiple modalities with advanced model architectures, data, and wiser use of different feature extraction, classification, and preprocessing techniques, assisting the task with improved performance and accuracy. In this work, we presented a study on multimodal deception detection using real-life data from 100 subjects belonging to the Indian population and spanning seven modalities exploiting the cognitive aspect of deception detection. We experimented with a novel primary multimodal deception detection dataset, CogniModal-D, consisting of seven distinct modalities: EEG, ECG, EOG, Gaze, GSR, audio, and video. Experiments are performed on the proposed interrogation-based mock crime and best friend paradigms that express social relationships designed for deceit detection based on multiple modalities. Classifiers based on the individual or combined sets of verbal and nonverbal features exploiting cognitive and behavioral aspects of deception are designed. It is observed that multimodal score-level-based fusion combination detects deceptive samples with an accuracy of 75% in the controlled mock crime interrogations and 79% in the best friend scenario involving storytelling about social relationships compared to their unimodal counterparts in distinguishing deceptive and truthful samples. Our analysis of verbal and non-verbal behavioral cues occurring in deceptive and truthful samples brought insights into the discriminant cues that play a significant role in deception detection. These insights will lead us towards building more advanced deception detection systems and algorithms involving multiple modalities to create a comprehensive and all-inclusive multimodal deception detection system emphasizing multimodal fusion, which provides improved performance and better results in comparison to their unimodal counterparts. The fusion of modalities improves task performance and enlists the impact and contributions of individual modalities. Modalities show complementary and assistive behavior, boosting the overall accuracy of the task. In the future, increasing the number of modalities and samples with advanced architectures will improve task performance, further enhancing multimodal deception detection capabilities. The inclusion of large language models in this arena builds context and offers a promising solution further. The research aids in the development of automated deception detection techniques exploring cognitive and behavioral aspects. Experiments with more data collected from real, everyday, and diverse conditions would produce more robust solutions and raise results and techniques to a level of potential industrialization and commercialization. We hope that our dataset and analysis will be aided by multiple modalities and that the findings will open diverse avenues of research that will assist in paving the road toward advanced multimodal deception detection algorithms and systems for multimodal cognitive behavior analysis in the future.

Data availability

The dataset experimented with in this work is the first comprehensive multimodal deception detection dataset for cognitive behavior analysis targeting the Indian population with over seven modalities and 100 subjects. The data will be made available to the research community upon request to the corresponding authors.

Received: 28 December 2024; Accepted: 27 February 2025
Published online: 15 March 2025

References

1. D'Ulizia, A., D'Andrea, A., Grifoni, P. & Ferri, F. Detecting deceptive behaviours through facial cues from videos: A systematic review. *Appl. Sci.* **13** (16). https://doi.org/10.3390/app13169188 (2023).
2. Wu, Z., Singh, B., Davis, L. S. & Subrahmanian, V. S. Deception detection in videos, *32nd AAAI Conf. Artif. Intell. AAAI 2018*, pp. 1695–1702, (2018). https://doi.org/10.1609/aaai.v32i1.11502
3. Krishnamurthy, G., Majumder, N., Poria, S. & Cambria, E. A deep learning approach for multimodal deception detection. *Lect Notes Comput. Sci. (including Subser. Lect Notes Artif. Intell. Lect Notes Bioinformatics)*. **13396**, 87–96. https://doi.org/10.1007/978-3-031-23793-5_8 (2023). LNCS.
4. Bhatt, P. et al. Machine learning for cognitive behavioral analysis: datasets, methods, paradigms, and research directions. *Brain Inf.* **10** (1). https://doi.org/10.1186/s40708-023-00196-6 (2023).
5. Constancio, A. S. et al. *Deception detection with machine learning: A systematic review and statistical analysis*, vol. 18, no. 2 February. (2023).
6. Bicer, B. & Dibeklioglu, H. Automatic deceit detection through multimodal analysis of High-Stake Court-Trials. *IEEE Trans. Affect. Comput.* **15** (1), 342–356. https://doi.org/10.1109/TAFFC.2023.3322331 (2024).
7. Derakhshan, A., Mikaeili, M., Gedeon, T. & Nasrabadi, A. M. Gender-based multimodal deception detection. *Multimodal Technol. Interact.* **4** (2), 1–14. https://doi.org/10.3390/mti4020025 (2020).
8. Pérez-Rosas, V., Mihalcea, R., Narvaez, A. & Burzo, M. A multimodal dataset for deception detection, *Proc. 9th Int. Conf. Lang. Resour. Eval. Lr.* pp. 3118–3122, 2014. (2014).
9. Prome, S. A., Raghavan, N. A., Islam, M. R., Asirvatham, D. & Jegathesan, A. J. Deception detection using machine learning (ML) and deep learning (DL) techniques: A systematic review, *Nat. Lang. Process. J.*, vol. 6, no. October p. 100057, 2024, (2023). https://doi.org/10.1016/j.nlp.2024.100057
10. Suchotzki, K. & Gamer, M. Detecting deception with artificial intelligence: promises and perils, *Trends Cogn. Sci.*, vol. xx, no. Xx, pp. 1–3, (2024). https://doi.org/10.1016/j.tics.2024.04.002
11. Wu, Y. C., Liu, Y. C. & Huang, R. Y. The use of artificial intelligence in interrogation: Lies and truth. *IAES Int. J. Robot Autom.* **12** (4), 332–340. https://doi.org/10.11591/ijra.v12i4.pp332-340 (2023).
12. Diana, B. et al. Multimodal deception detection: A t-pattern approach, *WMDD 2015 - Proc. ACM Work. Multimodal Decept. Detect. Co-located with ICMI 2015*, pp. 21–28, (2015). https://doi.org/10.1145/2823465.2823466
13. Dinges, L. et al. Automated Deception Detection from Videos: Using End-to-End Learning Based High-Level Features and Classification Approaches,., [Online]. (2023). Available: http://arxiv.org/abs/2307.06625

14. Belavadi, V. et al. MultiModal Deception Detection: Accuracy, Applicability, and Generalizability*, *Proc. – 2020 2nd IEEE Int. Conf. Trust. Priv. Secure. Intell. Syst. Appl. TPS-ISA 2020*, pp. 99–106, (2020). https://doi.org/10.1109/TPS-ISA50397.2020.00023
15. Gupta, V. et al. Bag-of-lies: A multimodal dataset for deception detection, *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, vol. 2019-June, no. ii, pp. 83–90, (2019). https://doi.org/10.1109/CVPRW.2019.00016
16. Lloyd, E. P. et al. Diversity DeMiami uniception detection database. *Behav. Res. Methods*. **51** (1), 429–439. https://doi.org/10.3758/s13428-018-1061-4 (2019).
17. Sen, M. U. et al. Multimodal deception detection using Real-Life trial data. *IEEE Trans. Affect. Comput.* **13** (1), 306–319. https://doi.org/10.1109/TAFFC.2020.3015684 (2022).
18. Kamboj, M., Hessler, C., Asnani, P., Riani, K. & Abouelenien, M. c, *IEEE Multimed.*, vol. 28, no. 1, pp. 94–102, (2021). https://doi.org/10.1109/MMUL.2020.3048044
19. Zhang, J., Levitan, S. I. & Hirschberg, J. Was It You Who Stole 500 Rubles? - The Multimodal Deception Detection, *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020-Octob, pp. 359–363, (2020). https://doi.org/10.21437/Interspeech.2020-2320
20. Gogate, M., Adeel, A. & Hussain, A. Deep learning driven multimodal fusion for automated deception detection, *2017 IEEE Symp. Ser. Comput. Intell. SSCI 2017 - Proc.*, vol. 2018-Janua, pp. 1–6, (2018). https://doi.org/10.1109/SSCI.2017.8285382
21. Ulizia, A. D., Andrea, A. D., Grifoni, P. & Ferri, F. Analysis, Evaluation, and Future Directions on Multimodal Deception Detection, (2024).
22. Rill-Garcia, R., Escalante, H. J., Villasenor-Pineda, L. & Reyes-Meza, V. High-level features for multimodal deception detection in videos, *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, vol. 2019-June, pp. 1565–1573, (2019). https://doi.org/10.1109/CVPRW.2019.00198
23. Mathur, L. & Matarić, M. J. Introducing Representations of Facial Affect in Automated Multimodal Deception Detection, *ICMI 2020 - Proc. 2020 Int. Conf. Multimodal Interact.*, pp. 305–314, (2020). https://doi.org/10.1145/3382507.3418864
24. UNSUPERVISED AUDIO-VISUAL SUBSPACE ALIGNMENT FOR HIGH-STAKES & DECEPTION DETECTION Leena Mathur and Maja J Matari' c Department of Computer. Science University of Southern California, Los Angeles, CA, pp. 2255–2259, (2021).
25. Karnati, M., Seal, A., Yazidi, A. & Krejcar, O. LieNet: A deep Convolution neural network framework for detecting deception. *IEEE Trans. Cogn. Dev. Syst.* **14** (3), 971–984. https://doi.org/10.1109/TCDS.2021.3086011 (2022).
26. Guo, X. et al. Benchmarking Cross-Domain Audio-Visual Deception Detection, pp. 1–10, [Online]. (2024). Available: http://arxiv.org/abs/2405.06995
27. Touma, L., Horani, M. A., Tailouni, M., Dahabiah, A. & Jallad, K. A. Voting-based Multimodal Automatic Deception Detection. arXiv preprint arXiv:2307.07516. (2023).
28. Karimi, H. Interpretable multimodal deception detection in videos. *ICMI 2018 - Proc. 2018 Int. Conf. Multimodal Interact.* 511–515. https://doi.org/10.1145/3242969.3264967 (2018).
29. Fernandes, S. V. & Ullah, M. S. A comprehensive review on features extraction and features matching techniques for deception detection. *IEEE Access.* **10**, 28233–28246. https://doi.org/10.1109/ACCESS.2022.3157821 (2022).
30. Fathima Bareeda, E. P., Shajee Mohan, B. S. & Ahammed Muneer, K. V. Lie detection using speech processing techniques. *J. Phys. Conf. Ser.* **1921** (1). https://doi.org/10.1088/1742-6596/1921/1/012028 (2021).
31. Yang, J. T., Liu, G. M. & Huang, S. C. H. Multimodal Deception Detection in Videos via Analyzing Emotional State-based Feature, [Online]. (2021). Available: http://arxiv.org/abs/2104.08373
32. Soldner, F., Pérez-Rosas, V. & Mihalcea, R. Box of Lies*: Multimodal deception detection in dialogues, *NAACL HLT 2019–2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 1768–1777, (2019). https://doi.org/10.18653/v1/n19-1175
33. Zhang, J., Levitan, S. I. & Hirschberg, J. Multimodal Deception Detection Using Automatically Extracted Acoustic, Visual, and Lexical Features. In INTERSPEECH (pp. 359–363) (2020), October.
34. Abouelenien, M., Pérez-Rosas, V., Zhao, B., Mihalcea, R. & Burzo, M. Gender-based multimodal deception detection, *Proc. ACM Symp. Appl. Comput.*, vol. Part F1280, pp. 137–144, (2017). https://doi.org/10.1145/3019612.3019644
35. Loy, J. E., Rohde, H., Corley, M. & Identifying, *J. Cogn.*, **1**, 1, 1–21, doi: https://doi.org/10.5334/joc.46. (2018).
36. Brennen, T. & Magnussen, S. Lie detection: what works?? *Curr. Dir. Psychol. Sci.* **32** (5), 395–401. https://doi.org/10.1177/09637214231173095 (2023).
37. Saini, R. & Rani, P. LDM: A Systematic Review on Lie detection Methodologies, no. December, pp. 1–18, (2022). https://doi.org/10.20944/preprints202212.0443.v1
38. Gupta, S., Kumar, P. & Tekchandani, R. A multimodal facial cues based engagement detection system in e-learning context using deep learning approach. *Multimed Tools Appl.* **82** (18), 28589–28615. https://doi.org/10.1007/s11042-023-14392-3 (2023).
39. Victor, D., Zizhao, C. & E. W. W., and Enhancing deception detection with exclusive visual features using deep learning. *Int. J. Perform. Eng.* **19**, 547. https://doi.org/10.23940/ijpe.23.08.p7.547558 (2023).
40. Camara, M. K., Postal, A., Maul, T. H. & Paetzold, G. Can lies be faked? Comparing low-stakes and high-stakes deception video datasets from a Machine Learning perspective, [Online]. (2022). Available: http://arxiv.org/abs/2211.13035
41. Gallardo-Antolín, A. & Montero, J. M. Detecting deception from gaze and speech using a multimodal attention LSTM-based framework. *Appl. Sci.* **11** (14). https://doi.org/10.3390/app11146393 (2021).
42. Li, P., Abouelenien, M. & Mihalcea, R. Deception Detection from Linguistic and Physiological Data Streams Using Bimodal Convolutional Neural Networks, [Online]. (2023). Available: http://arxiv.org/abs/2311.10944
43. Guo, X. et al. Audio-Visual Deception Detection: DOLOS Dataset and Parameter-Efficient Crossmodal Learning, [Online]. (2023). Available: http://arxiv.org/abs/2303.12745
44. Ding, M., Zhao, A., Lu, Z., Xiang, T. & Wen, J. R. Face-focused cross-stream network for deception detection in videos, *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, no. 2, pp. 7794–7803, (2019). https://doi.org/10.1109/CVPR.2019.00799
45. Abdulridha, F. & Albaker, B. M. A review on Non-Invasive multimodal approaches to detect deception based on machine learning techniques, pp. 619–633, (2024).
46. Alreshidi, I., Moulitsas, I. & Jenkins, K. W. Multimodal approach for pilot mental state detection based on EEG. *Sensors* **23** (17). https://doi.org/10.3390/s23177350 (2023).
47. Wang, X., Ge, S., Chen, X. & Walls, B. L. Deception detection with nonverbal behaviors from silence and speech time, *27th Annu. Am. Conf. Inf. Syst. AMCIS* no. August, 2021. (2021).
48. Abdul, Z. K. & Al-Talabani, A. K. Mel Frequency Cepstral Coefficient and its Applications: A Review, *IEEE Access*, vol. 10, no. November, pp. 122136–122158, (2022). https://doi.org/10.1109/ACCESS.2022.3223444

Acknowledgements

The authors acknowledge the support and funding from the Ministry of Electronics and Information Technology (MeitY), Government of India, under the grant sanction ID: 4(13)/2021-ITEA.

Author contributions

GJ, VT, AD -collected the primary data. AD, BS, AK, AS, KK, SK, YW, HM, SG, AS - preprocessing and model development. GJ, VT, AD, RW- Analysis and interpretation of results. GJ, RW- Writing the first draft and re-writing. RW, KK, PJ, NKJ - Funding acquisition review and editing. RW, KK - Supervision and mentoring. All authors read and approved the final manuscript.

Funding

Open access funding provided by Symbiosis International (Deemed University).

The work is funded by the Ministry of Electronics and Information Technology MeitY, Government of India (Project Sanction ID: 4(13)/2021-ITEA).

Declarations

Ethics approval and consent to participate

All methods were carried out in accordance with relevant guidelines and regulations as per the ethical approval obtained from Symbiosis International Deemed University's (SIU) Institutional Ethics Committee with approval number SIU/IEC/517. The study and procedure along with its use for research purpose was explained to all participants and informed consent was obtained from all subjects.

Competing interests

The authors declare no competing interests.

Consent

All subjects were above 18 years of age, and informed consent was obtained from all subjects.

Additional information

Correspondence and requests for materials should be addressed to R.W. or K.K.