



Exploring facial cues: automated deception detection using artificial intelligence

Laslo Dinges¹ · Marc-André Fiedler¹ · Ayoub Al-Hamadi¹ · Thorsten Hempel¹ · Ahmed Abdelrahman¹ · Joachim Weimann² · Dmitri Bershadskyy² · Johann Steiner³

Received: 5 July 2023 / Accepted: 12 April 2024
© The Author(s) 2024

Abstract

Deception detection is an interdisciplinary field attracting researchers from psychology, criminology, computer science, and economics. Automated deception detection presents unique challenges compared to traditional polygraph tests, but also offers novel economic applications. In this spirit, we propose an approach combining deep learning with discriminative models for deception detection. Therefore, we train CNNs for the facial modalities of gaze, head pose, and facial expressions, allowing us to compute facial cues. Due to the very limited availability of training data for deception, we utilize early fusion on the CNN outputs to perform deception classification. We evaluate our approach on five datasets, including four well-known publicly available datasets and a new economically motivated rolling dice experiment. Results reveal performance differences among modalities, with facial expressions outperforming gaze and head pose overall. Combining multiple modalities and feature selection consistently enhances detection performance. The observed variations in expressed features across datasets with different contexts affirm the importance of scenario-specific training data for effective deception detection, further indicating the influence of context on deceptive behavior. Cross-dataset experiments reinforce these findings. Notably, low-stake datasets, including the rolling dice Experiment, present more challenges for deception detection compared to the high-stake Real-Life trials dataset. Nevertheless, various evaluation measures show deception detection performance surpassing chance levels. Our proposed approach and comprehensive evaluation highlight the challenges and potential of automating deception detection from facial cues, offering promise for future research.

Keywords Deception detection · Facial cues · Deep learning · Rolling dice

1 Introduction

The study of deception detection is a complex and multi-disciplinary field that has gathered significant research attention. Several kinds of techniques have been proposed to detect signs of deception, including traditional methods such as polygraph tests and modern approaches utilizing computer vision, natural language processing, and machine learning [1]. Although the primary applications of deception detection techniques have traditionally been in the field of criminalistics, with a focus on interrogation, recent developments have shown that it has also become a significant concern in other domains such as border security [2], where automatized contact-free or even online approaches are preferred over traditional polygraph tests, which

are too complicated and costly for mass screening. However, using detection approaches in such a way—that might lead to conviction or the restriction of the freedom to travel—is delicate due to possible unreliable or discriminating decisions.

The significance of digitization and machine learning-based approaches also has now become a critical issue in the field of experimental economics [3–5]. In this regard, also automatized deception detection is of interest and, in contrast to the aforementioned domains, it can be used in a less invasive way. Salespeople who misrepresent the quality of their products can cause significant economic harm. Such fraudulent activities not only affect the buyers but also harm salespeople who offer high-quality products but suffer losses due to customers' cautious purchasing behaviors as a result of bad experiences [6]. People often overestimate their ability to detect lies, when in reality their

Extended author information available on the last page of the article

accuracy is only slightly better than chance, and at the same time they underestimate their ability to tell lies. Hence, an automated deception detection system could be useful, for example, to assist customers during virtual sales meetings using video and audio signals from regular webcams. Although accuracy may be lower compared to professional polygraph tests (which are hardly feasible in this context), they can still provide a valuable contribution to society by warning customers of potential deception as greatly exaggerated product quality.

Also in healthcare, patient honesty is pivotal for effective treatment. Despite doctors' awareness of potential deception, a balance of trust is essential. Admission of patients with factitious disorder or malingering presents a unique challenge in hospitals, involving feigned, exaggerated, or fabricated medical or psychiatric symptoms. Effective managing deception in such cases is crucial for accurate diagnosis and effective treatment [7].

In the following, we focus on the lie—that can be a wrong description of a photo or object that was shown to subjects or a false statement in a legal context—while other forms of deception as concealment, evasion, gaslighting, or trick scam are beyond the scope of this work. However, because the used databases include more than binary responses, they may also include half-truths, misleading, exaggeration, or manipulation. Furthermore, also the investigated signs are not restricted to a single modality, since many features such as facial expression, voice, transcription, vital parameters, and body language might be cues for deceptions as lies. As a matter of fact, we propose a multi-modal approach to evaluate some of the most promising features of contact-free deception detection—namely facial cues—on the four publicly available deception databases. Afterward, we employ this approach to evaluate a classical experiment known as the dice-rolling experiment [8].

2 Related works

In the following, we give an overview of related work in the area of general deception detection as well as from the point of view of AI-based automation.

2.1 Polygraph tests

Polygraph testing is one of the most widely used techniques for detecting deception. Polygraph tests measure physiological responses such as heart rate, blood pressure, and skin conductivity to determine whether a person is lying. Despite their widespread use, the accuracy of

polygraph tests has been a topic of debate in the scientific community [9].

There are several types of polygraph tests, such as the Control Question Technique (CQT) or the less widely used Concealed Information Test (CIT). These, however, do not differ in the kind of features that are used to detect deception, but in the kind of questions that are asked during interrogation. Unlike CIT, the CQT can be used even without critical information, however, CQT has been even more criticized than CIT due to its unethical testing conditions and highly questionable assumptions [9].

In general, polygraph tests are judged clearly less suitable by the scientific community than by the American Polygraph Association, for example, since deceptive answers will not necessarily produce unique cues or might be sensitive to cultural or other context [2, 10]. Another issue is that truth-tellers more often believe that their innocence is obvious (known as the illusion of transparency), which can make truth-tellers even less credible, for example, since they tend to react more aggressively than deceptive people if they feel they are not believed [11]. This is one of many reasons why traditional polygraph tests—but also automatized deception detection—will only indicate the possibility of deception in real-life scenarios, rather than providing a definitive identification. However, these methods are still clearly better than chance, while purely manual attempts are only around 2–4% more accurate [9, 12, 13]. This is true even for professionals who routinely engage in deception detection and can fall back on a variety of cues (behavior, administrative data, but also crime records). Vasu et al. found that one important reason is that even trained people have problems considering more than a single cue at a time [14]. Junger et al. found that manual strategies significantly improve online fraud detection if relying on pre-existing knowledge. (Fraud scheme is known.) However, it is noteworthy that many non-knowledge-based strategies not only prove ineffective, but also carry the potential to escalate the risk of falling victim to fraud [15]. On the other hand, using knowledge-based strategies is not always possible, e.g., in sales meetings where the quality of a not-yet-rated product is highly exaggerated (resulting in information asymmetry), and only behavioral information is available.

2.2 Automatized deception detection

In the following, we review some key works of contact-free, automatized uni- and multi-modal deception detection.

While polygraph tests are criticized but still relatively reliable due to contact measurements such as skin conductance and, even more important, due to an expert that

interrogates the candidates, approaches of automatized deception detection are typically limited to contact-free video or audio data modalities and available databases do not involve any baselines as control questions. Since available databases are also not comprehensive, automatized detection of deception is even more challenging. However, since they are simple to use and involve much lower cost, new areas of applications are enabled, for example, as an online service. Typical contact-free modalities which are used for automatized deception detection are micro-expressions, macro-expressions, thermal images, gaze, gestures, voice features (tone and pitch), or the transcription of what was said [1, 16].

2.2.1 Gaze and head pose

While body poses and gestures can be effective features for detecting deception in certain contexts [17], it may not be the case in all situations. For instance, in our targeted application scenario and most of the available deception databases, subjects are seated and do not exhibit significant changes in their body language or hand-gestures that can be used for deception detection purposes. However, head pose and gaze direction can also serve as potential indicators of the user's state [18].

Head pose Head pose estimation from a single image is a crucial task for various applications, including driver assistance, human–robot interaction, and even pain detection [19]. Apart from that, people often avoid eye contact when they feel guilty (for example, after being untruthful), which makes tracking changes in head pose and gaze direction valuable for lie detection purposes [18]. Two types of methods are commonly used for this task: landmark-based and landmark-free. Landmark-based methods require accurate landmark detection for accurate 3D head pose estimation [20], while landmark-free methods estimate head pose directly using deep neural networks to formulate orientation prediction as an appearance-based task [21–23]. Landmark-based methods can be affected by occlusion and extreme rotation, which is why landmark-free approaches perform better on comprehensive but challenging datasets.

Gaze Gaze estimation can be achieved by conventional regression-based methods; however, recently, approaches that are based on Convolutional Neural Networks (CNNs) are favored. CNNs have shown promise in modeling the

nonlinear mapping function between images and gaze, with researchers proposing different architectures that take into account factors such as eye images, full-face images, and head pose [24]. Other approaches include combining statistical models with deep learning and using temporal models such as Long Short-Term Memory networks (LSTMs) [25, 26]. Another work is AGE-Net, which uses two parallel networks for each eye image and an attention-based network to generate a weight feature vector, with outputs multiplied and refined with the output of VGG as CNN of the face images [27].

In the context of deception detection, several studies have utilized gaze features. Kumar et al. extracted statistical features from the gaze of a group of individuals playing the game Resistance [28]. They found that deceivers change their focus less often than non-deceivers, indicating less engagement in the game. Beyond gaze direction, Pasquali et al. report that the mean pupil dilation is a good indication of deception of children when playing a game. They received the promising lie-classification F1-Score of 56.5% for Random Forest (RF), respectively, 67.7% for Support Vector Machine (SVM) [29]. However, they did not compare this feature with any others, such as gaze direction or blinking. Furthermore, their approach requires additional, not contact-free hardware (Tobii Pro Glasses 2 eye tracker) to ensure accurate measurement of pupil dilation. Gupta et al. use the PyGaze library to extract fixations, eye blinks and pupil size as features. They got 57.11% accuracy on the bag-of-lies database which is better than accuracy for video (Local Binary Pattern), audio (frequency features), and EEG modality. Combining features from all modalities, 66% accuracy was achieved and 60% using all contact-free modalities [12].

2.2.2 Facial expressions

Facial expressions are a common way to detect basic emotions, which are typically conveyed through specific sets of facial movements known as Action Units (AUs) [11]. Constancio et al. found that, in general, emotional features are crucial for deception detection, as the act of deceiving triggers emotional states, and emotions are primarily communicated through facial expressions [30]. These expressions can be categorized into micro- and macro-expressions, which we will describe in the next sections.

Micro-expressions Micro-expressions are brief, involuntary facial expressions that occur when a person tries to conceal their true feelings or intentions [31]. They last less than half a second, making them difficult to detect. However, thanks to technological advances, it is now possible to automatically detect and analyze micro-expressions, so they could be used for lie detection [32].

Some studies have shown that (manually detected) micro-expressions (ME) can be used for accurate detection of deception, as they reveal emotions which a person tries to hide [33, 34]. However, it should be noted that micro-expressions do not always indicate deception, as they can also be caused by other factors such as stress or anxiety. Therefore, automatic deception detection systems that rely on micro-expressions should be used in conjunction with other methods to achieve the most accurate results [11]. According to Jordan et al., the practical use of ME for lie detection is even questionable, and it may not enhance the overall accuracy of the results [35]. One reason why using micro-expressions (ME) for lie detection is limited is that they occur in both genuine and deceptive reactions and occur far too infrequently and in ways that are too variable to be useful in detecting deception [11, 35]. Also, the infrequent occurrence reduces its quality as a feature for manual, computer-assisted or automatized lie detection. Hence, even with a model for ME detection or classification that generalizes well—so it could be used beyond controlled laboratory conditions and with 30–60 instead of 100–200 frames-per-second(fps)—applicability of ME in real-life deception detection scenarios would be still limited at the present time. Additionally, even training such a model for detecting and classifying ME within unseen data in the first place is challenging: The available databases (as CASME II or SAMM) are limited in sense of comprehensiveness, general applicable experimental design, and kind of covered Action-Units [36, 37].

Macro-expressions Macro-expressions are regular facial expressions described by the facial Action-Unit (AU) coding system [38]. AUs as Cheek Raiser (AU6) or Brow Lowerer (AU4) measure observable muscle activations in specific regions of the face and can be used to predict underlying basic emotions, pain or other secondary emotions from individual or sequences of images. In-the-wild, macro-expressions can sometimes result in AU of weak intensity, but will still last longer than micro-expressions. Since the reliable, unambiguous assignment to exactly one basic emotion or neutral is barely feasible for many in-the-wild samples, the regression of continuous values for

valence and arousal values is used instead or additionally, following the idea of the circumplex model [39].

Previous methods for facial expression recognition mixing deep learning with other methods, following a conventional pipeline of face recognition, landmark extraction, and regression of AUs, which are then interpreted by a classifier to derive expressed basic or secondary emotions as pain [40, 41]. In this regard, training several partial solutions on—in some cases small and under controlled conditions acquired—databases is problematic, as, for example, landmark extraction might fail when applied to challenging in-the-wild samples. Hence, similar to head pose detection, landmark-free approaches that estimate AUs [42], basic emotions or valence/arousal values directly from raw images are more effective on challenging datasets, provided there are enough available training samples [43].

In their study, Chang et al. [44] used a CNN to compute AUs, which were subsequently used to predict valence and arousal values. On the other hand, Khorrami et al. [45] used a holistic strategy to directly predict valence and arousal values from (normalized) face images. AUs were implicitly extracted using a CNN.

Mollahosseini et al. introduced the comprehensive AffectNet database, which covers labels for basic emotion as well as continuous labels for valence and arousal [46]. On this dataset, they conducted a comparison between traditional approaches, such as HoG features and Support Vector Regression (SVR), and deep learning methods, which demonstrated superior performance. Zhang et al. took a similar approach, where they used a pre-trained network to make predictions about emotions and personalities [47]. Li et al. trained a bidirectional RNN that incorporated future images for prediction [48].

In the study of Chu et al., they proposed to extract both spatial representations from the data using a CNN and temporal representations using a LSTMs [49]. Their approach merged the results of the CNN and LSTMs models, resulting in better performance in predicting AUs. However, no further emotion classification was performed in their study.

2.2.3 Other modalities

As previously mentioned, hand gestures may not be appropriate in situations where individuals infrequently utilize gestures. Nevertheless, Avola et al., in their study [50], which builds upon their earlier work in 2020 [17], demonstrated an impressive accuracy rate of 95%.

However, results are only reported for a selected subset of the Real-life trial (RL) (77 samples that show suitable gestures) and are therefore hard to compare. They employed a set of manually engineered features, including measurements such as palm-fingertip distance, derived from the 42 OpenPose keypoints associated with hand joints. Their analysis revealed that the speed distribution of the index finger significantly differs between honest and dishonest subjects, with dishonest subjects exhibiting an almost negligible speed (close to 0) in contrast to the 175 speed units exhibited by honest subjects. These carefully crafted features were subsequently used for training an LSTM model.

A special kind of feature can be extracted from thermal cameras. Satpathi et al. [51] derive the normalized blood flow rate from regions of interest (forehead and periorbital region) of such thermal data from 10 subjects. They found that blood flow raising steeper for subjects which lie. However, as expensive, specialized equipment, thermal cameras are barely suitable for online applications as virtual sales meetings. Hence, it is of essential interest if such features can also be predicted using plain RGB-cameras. This would still imply some requirements—as good video quality and low variations of light, face pose, etc.—but is indeed possible for vital parameters as heart or respiratory rate [52].

When it comes to transcriptions, it is worth noting that individuals who engage in deception tend to be vaguer, containing less self-references and more negative and cognitive mechanism words [53, 54]. While providing specific details can enhance credibility, it also increases the risk of being disproved. This approach may be effective in a legal context or similar situations; however, it might be less useful if a deceiver just has to describe, for example, a product in one or two sentences without any interrogation. Ioannidis et al. give a detailed mathematical description of a statistical model for transcription-based lie detection for legal context [53]. However, this work is primarily theoretical and does not provide any concrete results that can be compared with other methods on a specific database.

IBorderCtrl is a tool that can potentially be used for mass screening in order to detect traveler that might be suspicious and need to be checked manually by border patrol. Sanchez et al. critically analyzed the limited public available information [2]. They found that the tool uses 38 features, or channels, such as ‘left eye blink,’ ‘increase in face redness’ or head movement directions. The system was trained on a small, lab condition database where 32 actors play the role of truthful or deceptive travelers, answering 13 questions (such as ‘Are there any items from the lists of prohibited items in your case?’). For deception

detection, an accuracy of 73.66% is reported and 75.55% for detecting truthfulness. However, due to the very limited number of (acted) training samples, Sanchez et al. doubt that the current system is suitable to detect deception in border patrol context without major drawbacks.

Results on the common available deception databases shall be discussed in the next section.

3 Available deception databases

In the following, we take a closer look at the four publicly available deception detection datasets, three of them have low-stake and one has high-stake context. In general, it is observed that labeled deception datasets from real-life data and languages other than English are scarce [30].

3.1 Bag-of-lies

Bag-of-lies (BgL) is a multimodal dataset designed for deception detection using various modalities, including video, audio, and, unlike other datasets, EEG data (for a subset of the database) [12]. The dataset aims to explore the cognitive aspect of deception and combines it with vision, providing a realistic scenario for collecting data. BgL contains 35 unique subjects, providing 325 annotated samples with an even distribution of truth and lies. The goal of this dataset is to facilitate the development of better deception detection algorithms that are more relevant to real-world scenarios. The samples are acquired as follows: A photo is displayed on a screen in front of a participant, who is then asked to describe it. The participant can freely choose whether to describe the photo truthfully or with deception. A motivation for a deceptive answer, as a higher reward, is not given.

Gupta et al. achieve 66.17 % Accuracy on a subset of BgL using EEG, gaze, video, and audio features. On the full BgL they got 60.09 % (without EEG). Using single modalities, gaze (using an eye tracker system) gives the best accuracy of 61.7%/57.11% (on subset/full set) results, while 56.2%/55.26% were achieved using video features only.

3.2 Box-of-lies

Box-of-lies (BxL) is a dataset that is based on the game with the same name that is part of a late night TV show. In this show, the guest and host take turns describing an object truthful or deceptive (or on the opponent’s move guessing whether this description might be the truth) [55]. Besides linguistic and dialog features, also various nonverbal cues

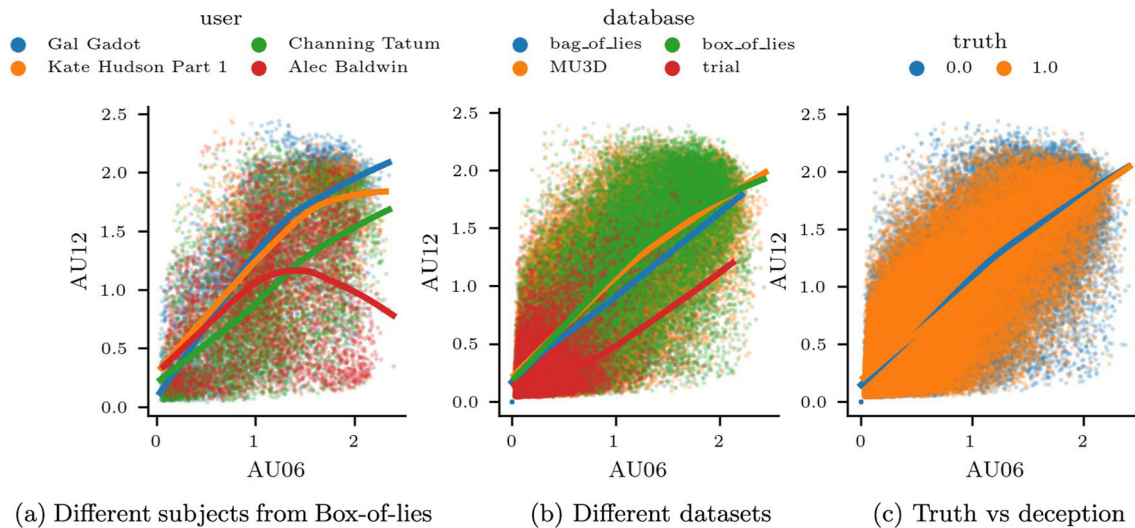


Fig. 1 Plot analysis the correlation of the Action Units AU12 (Lip corner puller) and AU06 (Cheek raiser) as indication of a false smile. Deviations can be observed between subjects, databases, however, not between true and lie-samples

are manually ground-truthed over time including features from gaze, eye, mouth, eyebrows, face and head.

Soldner et al. achieve 69% Accuracy (guests truthful/deceptive) training Random Forest on 60% of the samples using all the manually extracted features. Furthermore, the experiments showed that linguistic features were the most relevant (66% Accuracy), followed by non-verbal cues (61%) [55]. Furthermore, accuracy by Human, that is based on guess while watching the video samples, is just 55%. However, during an annotated scene—which is a time-slot in which the truthful or deceptive description should be given—cuts are often made, and the camera may shift to the 'opponent' or even both players. This as well as the fact that BOX is biased (65% deceptive answers) could be an obstacle using automatized feature extraction. Apart from that, most guest are actors, which could makes it difficult to transfer any results to real-world scenarios.

3.2.1 Preliminary investigation concerning fake smiles

As a first preliminary investigation, we tried to detect fake smiles using the measured intensities for the AUs AU12 (Lip Corner Pulled) and AU06 (Cheek Raiser). We expected that AU06 would be low compared to AU12 in case of faked smiles, since they should not involve the eye areas to the same extent. However, as one can see in Fig. 1c), the AUs correlate well for both truthful and deceptive samples. Nevertheless, there is a clear difference if we compare the four databases: the high-stake RL dataset, where the subjects were on trial, shows the lowest values for AU12 compared to the other datasets where

AU06 and AU12 show a better correlation (see Fig. 1b)). Furthermore, there is a significant individual influence, as shown in Fig. 1a).

3.3 Miami University deception detection database

The Miami University Deception Detection Database (MU3D) is a free resource that includes 320 videos of individuals telling truths and lies. The videos feature 80 different targets of different ethnic backgrounds. Each target produced four videos, covering positive and negative truths and lies, which results in a fully crossed dataset to investigate research questions as 'are positive lies more difficult to detect than negative lies?' or how target features as age influence impact deception detection [56]. The videos were transcribed and evaluated by naive raters, and descriptive analyses of the video characteristics and subjective ratings were provided. The MU3D offers standardized stimuli that can enhance replication among labs, promote the use of signal detection analyses, and facilitate research on the interactive effects of race and gender in deception detection.

3.3.1 Analysis of MU3D

A significant feature of MU3D, compared to the other datasets, is the valence ground truth, which is the positive or negative context: telling truth or lies about a person the subject likes or dislikes [56]. In the original paper, distributions are provided only for individual features, without

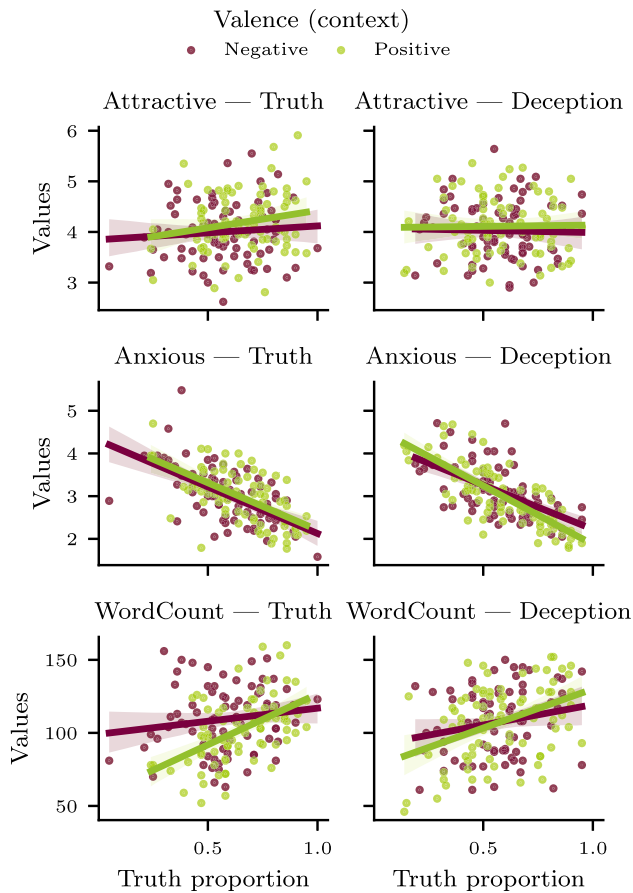


Fig. 2 Three manual features above the truth proportion, separated into positive (green) and negative (red) valence, and deceptive (left) and truthful (right) responses. Each subplot is a scatter plot with linear regression fit where each circle represents a single sample video. The truth proportion is the mean value of human raters' belief that the subject answers truthfully in a sample video (colour figure online)

exploring their correlation. Hence, we analyze the truth proportion (guessed by human raters) in relation to some features (which are part of the ground truth) and compare the results for positive and negative valence and for true and deceptive answers, as one can see in Fig. 2.

We discovered that as the level of anxiety increased among the subjects, the likelihood of the observers judging them as truthful decreased. However, we found no significant influence of the actual veracity (truth, deception) or the valence context in this regard. This shows firstly that human raters might overestimate the significance of noticed signs of anxiety as a cue for deception, and secondly that the influence of the valence context might not be crucial in case of low-stake scenarios. However, truth proportion is significantly more sensitive to the word count in case the proband speaks the truth and valence is positive (left bottom subplot of Fig. 2). Apart from that, the

(subjective) attractiveness has only a slight influence on truth proportion in all cases.

3.4 Real-life trials

Perez et al. note the importance of identifying deception in court trial data, given the high stakes involved [57]. In this vein, they introduce Real-life trial (RL)—a novel dataset consisting of videos from public court trials—and present a multimodal deception detection system that utilizes both verbal and non-verbal modalities to discriminate between truthful and deceptive statements made by defendants and witnesses.

The system achieves classification accuracies in the range of 60–75% using a set of manually extracted features and decision trees or RF for classification and leave-one-out cross-validation. This outperforms human capability to detect deception in trial hearings on raw data (62% accuracy). Their findings further indicated that facial features were the most useful for identifying deceit using Machine Learning (ML). Human observers, on the other hand, were most effective on audio-only or full-video, followed by audio, text and silent video (while agreement of 3 annotators was best for audio-only) [57].

3.4.1 Preliminary investigation on RL

We repeat the experiment using the given manual features as input for a SVM using repeated, random train-test splits with $k = 30$. To further analyze the dataset, we rank the feature according to their relevance and plot their occurrence for samples that are labeled as deceptive, respectively, truthful, as one can see in Fig. 3. Frowning, (eyebrow) raise, or lips down are one of the most relevant features, which are also among the best 6 features reported in [57]. Using all non-verbal features, we achieve an average accuracy of 76.2% for SVM, respectively, 74.9% for RF classification, which is comparable to Perez et al. which got 73.55% using RF and 68.59% using decision trees [57].

4 Methodology

In this section, we present our approach for deception detection. In Sect. 4.1, we describe our deep learning methods for generating comprehensive features from diverse modalities. Thereafter, we outline our proposed deception classification approaches in Sect. 4.2.

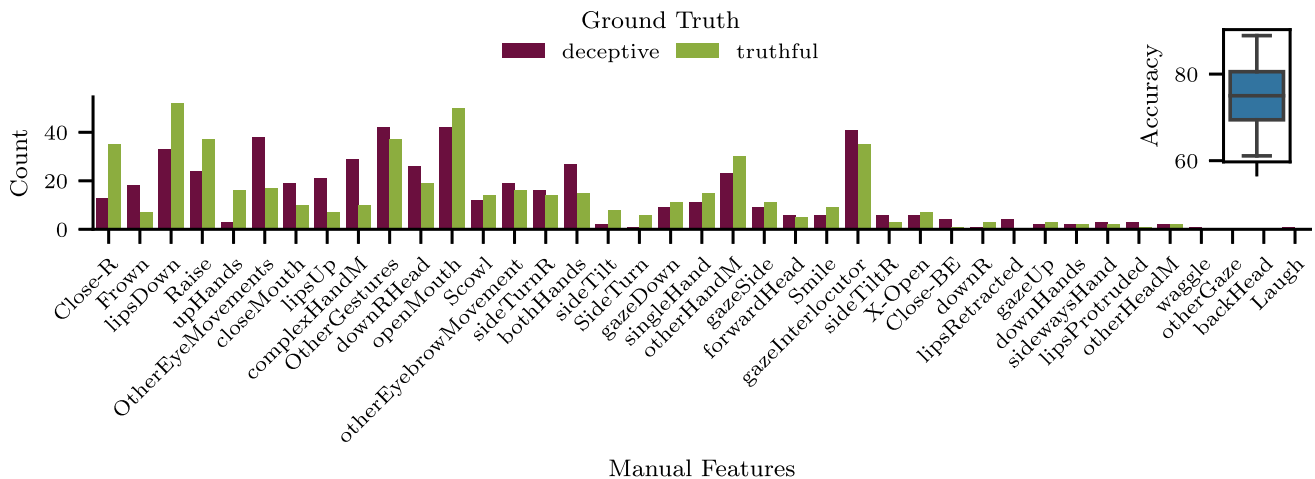


Fig. 3 Distribution of the given manually labeled features of the RL dataset for truth and lie samples. Features are ordered descending from left to right according to the achieved feature relevance based on cross-validation using SVM. On the upper right, there is a boxplot

including the median of the accuracy which is 77.7% while the average accuracy is 76.2% (using RF we get a median of 75.0% and an average of 74.9%)

4.1 Deep learning for visual modalities

In this section, we describe our research toward different modalities, namely head pose, gaze detection, and emotion recognition. Therefore, we train the CNN models on comprehensive datasets (except for AUs, where we used a ready to use model).

4.1.1 Head pose estimation

Current landmark-free methods often split up continuous rotation variables into bins for classification, leading to a loss of information. Additionally, using the popular Euler angle or quaternion representations for training neural networks for head pose estimation may not be ideal due to their discontinuity, known as *gimbal lock*, for Euler angles, and ambiguity for quaternions. To address this, we proposed a landmark-free head pose estimation method in [58] based on the rotation matrix representation, allowing for full pose regression without ambiguity. This simplifies the network by avoiding performance stabilizing measures used in other methods, such as discretization of rotation variables into a classification problem. Instead of predicting the entire nine-parameter rotation matrix, we regress a compressed 6D form. We then transform this into the 9D rotation matrix in a second step. Furthermore, we use the geodesic loss instead of the often used l_2 -norm to penalize the network in the training process, capturing the $SO(3)$ manifold geometry.

4.1.2 Gaze estimation

The majority of CNN-based gaze estimation models predict 3D gaze by regressing the (yaw, pitch) in spherical coordinates and use the mean-squared error for penalizing their networks. We propose a simple network architecture called L2CS-Net [59] based on the ResNet50 backbone and the combination of classification and regression losses. We propose to predict each gaze angle in a separate fully connected layer in order to utilize their independence and to capture the characteristic of each angle by detaching the shared features of the two angles from the last layer of the backbone. Instead of directly predicting continuous gaze angle values directly, we bin the gaze targets to combine classification and regression losses for a coarse-to-fine strategy that effectively promotes our network's gaze performance.

In addition, we utilize two separate losses, one for each gaze angle. Each loss consists of classification and regression components. We perform gaze classification by utilizing a softmax layer along with cross-entropy loss to obtain coarse gaze direction. On the other hand, we get fine-grained predictions by calculating the expectation of the gaze bin probabilities followed by a gaze regression loss.

4.1.3 Micro-expressions

In a previous experiment, we found that the available micro-expressions (ME) datasets were insufficient for proper training of robust, well generalizing models due to their limited size, uneven ME class distributions, and an strong overall bias toward certain classes [36]. To address

this issue, we combined two databases, using only the three most common ME classes. Although this improved intra-database results (approximately 80% accuracy), cross-database performance was still not better than random guess.

Considering the limitations of the available datasets, including a lower frame rate observed in all examined deception datasets, we currently do not recommend using ME for deception detection unless future datasets effectively address these limitations.

4.1.4 Macro-expressions

For macro-expressions, we use two models, one to detect AU intensities and one to predict expressed emotions directly.

Action-Units To predict AUs for the experiments of this article, we have used a model from Fan et al. that is based on the Resnet50 backbone and was trained on the BP4D database [60]. Heatmap regression and Semantic Correspondence Convolution were used to achieve reliable predictions of intensities for five different AUs. Fan et al. report good results which outperformed 6 other approaches (in average they got an ICC of 0.72 and MAE of 0.58). For future works, we plan to train our own models on EmotionNet [61], since it contains 12 AU classes and covers 100k manually labeled in-the-wild samples.

Multitask Emotion Prediction In addition to using AUs, we can also directly derive emotional states from facial expressions. While it is true that certain AUs can be used to derive emotions, annotating them is a time-consuming process which is why many comprehensive databases do not include AU labels. This is why we believe that investigating the use of both AUs and direct prediction of emotions has the potential to yield benefits regarding deception detection.

The comprehensive AffectNet database comprises approximately 500,000 samples, each with manual labels for basic emotion categories, as well as values for valence and arousal [46]. This allows for the simultaneous training of emotion classification models for neutral, happy, sad, surprise, fear, disgust, anger, and contempt (although contempt is considered a secondary emotion). Additionally, the database allows for the regression of intensities within the range of -1 to 1 for both valence and arousal.

To reduce the computational costs—which is crucial for our future experiments—we employ a multitask model to perform both classification and regression with a single network.

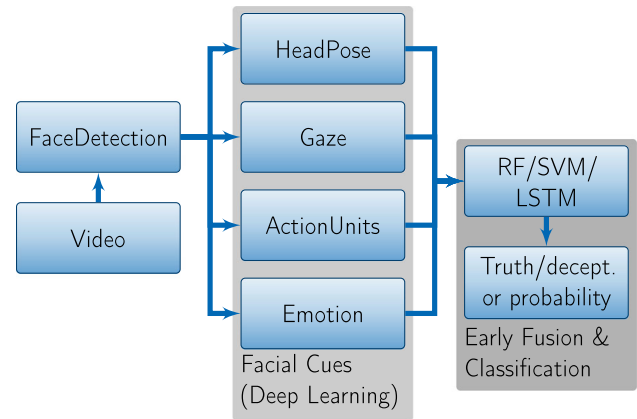


Fig. 4 Overview of our proposed approach for automatic deception detection using artificial intelligence

4.2 Deception detection

The above-described tasks as head pose estimation or emotion prediction were efficiently solved using deep learning due to the availability of at least one comprehensive dataset. However, when it comes to deception detection, all available datasets are clearly limited in respect of such a complex and highly context-sensitive task.

Transfer learning is not feasible in this case due to the absence of a typical pretraining domain. While macro-expressions exist as a feature of deception, they do not provide a comprehensive and suitable pretraining task. Without a single domain that adequately represents the target task, finding an appropriate base model through transfer learning becomes challenging. Instead, using several CNNs for multiple modalities/facial cues and late fusion using LSTMs or SVM/RF might be a better and more flexible strategy. Furthermore, signs of deception are typically ambiguous, very individual, and weakly expressed, especially in low-stake contexts. As a matter of fact, deep learning would require even more training data compared to tasks with measurable target values as (head pose) angles, but the datasets cover just a few hundred samples. Moreover, this approach exhibits a less pronounced black-box nature, which can be beneficial for certain applications. Hence, we use the above-discussed CNN models for head pose and gaze estimation, action unit regression, and classification of basic emotion as well as regression of valence & arousal (VA) values as input for early fusion to detect deception.

As shown in Fig. 4, we first apply face detection on the deception datasets using the RetinaFace model from Deng et al. [62]. Then, we use these face images as input for the

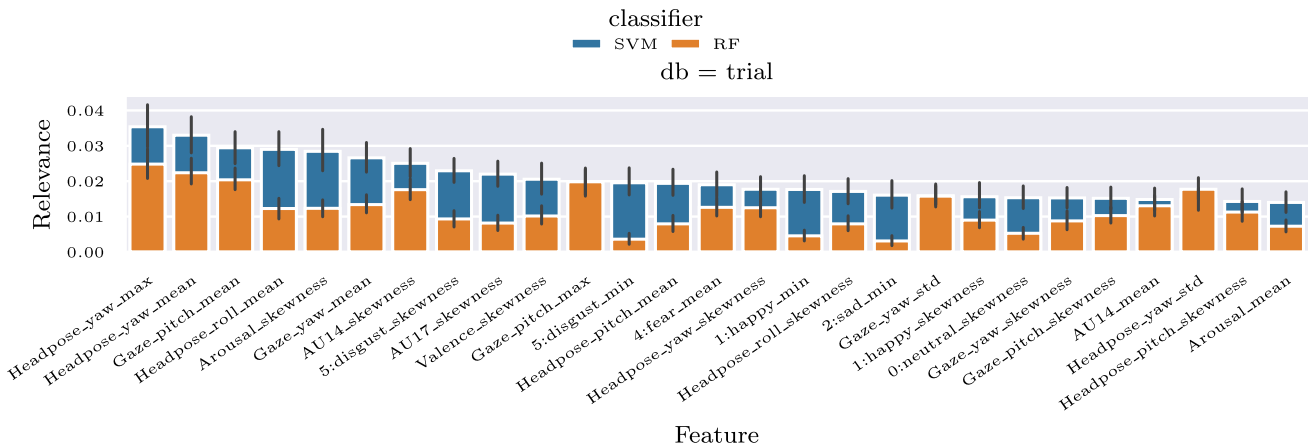


Fig. 5 Feature Ranking. We calculated feature relevance through feature permutation using traditional classifiers (SVM and Random Forest) for the RL dataset. The resulting list of features was ordered based on their relevance as determined by the SVM classifier, after

which we retained the top 25%. Notably, head pose and gaze features emerged as particularly relevant for the trial database. It is worth noting that the relevance of features can vary depending on the classifier used

CNNs, while their outputs are used as cues or in the context of machine learning (high-level) features. Through early fusion, those features are then employed to train LSTM models or discriminative classifiers, facilitating the detection of deception. (see Sect. 4.2.2 and 4.2.3).

4.2.1 Analysis of the facial cues

For each modality, we train a separate CNN model which has several outputs that are used as cues (base features), such as the yaw and pitch angle in case of gaze. As

modalities may not be suitable to the same extent for all datasets—depending on what meaningful patterns there are for these modalities in both the train and the test set—the significance of a single cue may depend on the data. Hence, we first compare the feature distributions of truthful and deceptive samples for the four datasets.

As shown in Fig. 6, the head pose yaw angle is a very interesting cue in the case of the RL dataset, since it is clearly higher in the case of deception. (Distribution for gaze yaw angle is quite similar.) This may be a result of the high-stake context since a guilty person might tend to avoid eye contact more.

In case of the valence feature, the distribution of deceptive and truthful samples is almost identical. However, it is obvious that is clearly differing for the different datasets. Unexpectedly—despite BgL being acquired under lab conditions and in a low-stakes context like MU3D—the valence distribution is predominantly within the negative

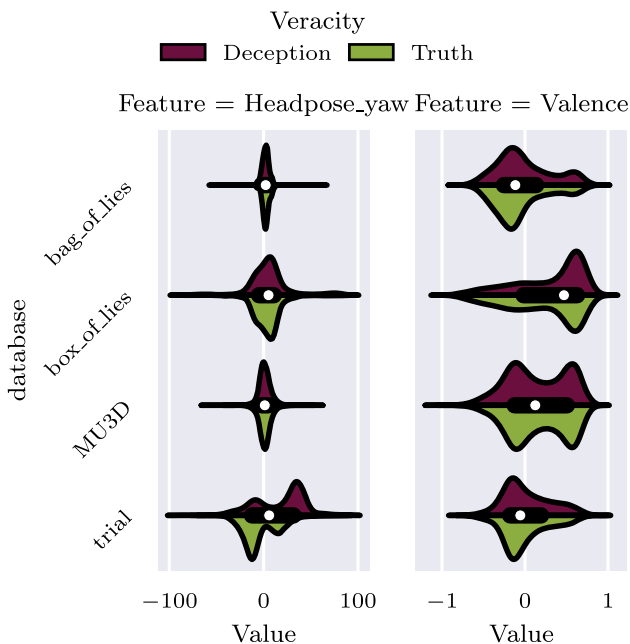


Fig. 6 Violinplots of key features extracted by deep learning approaches across four publicly available datasets, highlighting correlations with deceptive behavior

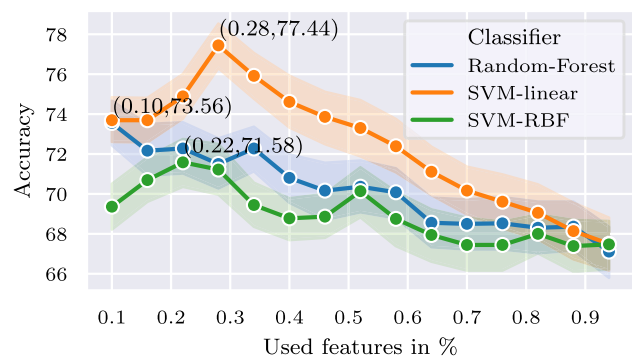


Fig. 7 Average and standard error (using cross-validation) for different fractions of the most relevant features for the discriminative classifiers

range, closely resembling that of RL datasets, which typically involve more serious contexts.

The distributions for all used features can be found in Appendix 7.

4.2.2 Early fusion using discriminative models

In the first step of our approach, we employ discriminative models such as SVM and RF to classify the video samples from deception databases as truthful or deceptive. To extract meaningful information from each base feature, we derive a range of statistical features such as mean and skewness. We then rank these features based on their relevance. Figure 5 shows the ranking of the top 25% of features for both linear SVM and RF for the RL dataset. As expected, yaw angle of gaze and head pose are among the most relevant features. For a comprehensive ranking of all databases, please refer to Appendix 7.

Feature selection In the next step, we aim to reduce the number of features used in our classification model. We have observed that retaining approximately 20–30% of the most relevant features yields good results and reduces computational costs. However, the optimal number of features depends on the dataset and the type of classifier used.

In Fig. 7, we present the performance of three classifiers on the RL dataset. Our results indicate that the best accuracy of 77% is achieved when using approximately 30% of the best features, while the use of all features results in worse performance for all tested classifiers. Nevertheless, the results obtained with all features are still better than those obtained by a trivial classifier or random guess, that both would have 50% accuracy.

For the BgL and BOX datasets, we have also found that retaining around 20%, respectively, 30% of the features yields the best performance, while for MU3D, we obtained the best results by keeping approximately 80% of the features.

Table 1 Comparison of the statistics of the four databases

Database	Samples	Subj	ϕ	Unbiased	ItW
BgL	325	35	9.2±4.8s	✓	×
BxL	25	25	154.9±45.2s	×	✓
MU3D	320	82	35.7±3.7s	✓	×
RL	118	56	24.9±14.2s	✓	✓
RDE	101	101	28.31±2.78	×	×

Here applies ϕ : average sample length in seconds and standard deviation, ×: No, ✓: Yes, Unbiased: ratio of truthful and deceptive samples is ≈ 0.5 , ItW: in-the-wild

We have also investigated the use of principal component analysis (PCA) for feature reduction. However, our results show that PCA does not perform well for our classification task. This may be due to the nonlinear relationships between the features and the outcome, which may be lost after PCA's linear combination, or the sensitivity of PCA to outliers in the data.

4.2.3 Early fusion using sequence-to-class approaches

Instead of computing statistical features for each sample video, time series analysis can be performed on the data using sequence-to-class approaches. We tried different Recurrent Neural Network (RNN) architectures including GRU, LSTMs and bidirectional LSTMs, where bidirectional LSTMs performed best.

4.3 Rolling dice experiment

The classical economically motivated Rolling dice Experiment (RDE) from Fischbacher et al. demonstrated that only 20% of all individuals who roll a dice without supervision will lie to the fullest extent possible (by claiming to have rolled the highest value when they actually did not), while 39% will be entirely truthful [8]. It is worth noting that the participants in this experiment knew that they were not monitored or recorded, so it was impossible to detect lies on an individual level.

In our study, both the participants and the results of the dice rolls are recorded. Participants are informed that the data will be labeled for analysis and research purposes, but that the experimenter will not have access to the records. This is done to ensure the integrity of the data, and only authorized individuals outside the economy research team will have access to the records.

To summarize our experimental setup:

1. We use a single standard six-sided dice, so the possible outcomes are 1, 2, 3, 4, 5, or 6.
2. Reward is 1 Euro times the outcome, except for a result of 6 which results in no reward.
3. We have 100 candidates each roll the dice once.
4. We analyze the results to determine the frequency of each outcome. We also record the claim and the actual dicing by webcam.

In the experimental section, we will compare the results of our modified rolling dice experiment with those of Fischbacher et al. and, furthermore, test our approach of automatic deception detection on the resulting dataset.

Table 2 Comparisons of our pose estimation with the state-of-the-art methods on the AFLW2000 and BIWI dataset

	Full Range ¹	AFLW2000				BIWI			
		Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
HopeNet ($\alpha = 1$) [21]	×	6.92	6.64	5.67	6.41	4.81	6.61	3.27	4.90
FSA-Net [63]	×	4.50	6.08	4.64	5.07	4.27	4.96	2.76	4.00
HPE [23]	×	4.80	6.18	4.87	5.28	3.12	5.18	4.57	4.29
QuatNet [22]	×	3.97	5.62	3.92	4.50	2.94	5.49	4.01	4.15
WHENet-V [64]	×	4.44	5.75	4.31	4.83	3.60	4.10	2.73	3.48
TriNet [65]	✓	4.04	5.77	4.20	4.67	4.11	4.76	3.05	3.97
FDN [66]	×	3.78	5.61	3.88	4.42	4.52	4.70	2.56	3.93
6DRepNet	✓	3.63	4.91	3.37	3.97	3.24	4.48	2.68	3.47

Bold values indicate the best result

All models are trained on the 300W-LP dataset. ¹ These methods allow full range predictions

Table 3 Comparison of mean angular error between our proposed model and SOTA methods on MPIIGaze dataset

Methods	MPIIFaceGaze
iTracker (AlexNet) [67]	5.6°
MeNets [68]	4.9°
FullFace (Spatial weights CNN) [69]	4.8°
Dilated-Net [70]	4.8°
RT-Gene (1 model) [71]	4.8°
GEDDNet [72]	4.5°
RT-Gene (4 ensemble) [71]	4.3°
Bayesian Approach [73]	4.3°
FAR-Net [74]	4.3°
CA-Net [75]	4.1°
AGE-Net [76]	4.09°
L2CS-Net L2CS-Net	3.96° 3.92°

5 Results and discussion

In this section, we present the results and engage in a comprehensive discussion to interpret the findings. In the following subsections, we will first discuss the results of the deep learning approaches for the facial cues. Subsequently, we will evaluate our approach of deception detection built on these results using the four available datasets. After examining our new RDE dataset, we will assess the proposed deception detection approach using this dataset and compare the outcomes with the other datasets.

5.1 Results for head pose estimation

We utilized PyTorch to implement our proposed network, selecting RepVGG (RepVGG-B1g2) as the backbone

model. RepVGG is designed as a multi-branch model for training and can be converted into a VGG-like architecture for deployment, resulting in the same accuracy but with a shorter inference time. We chose the RepVGG-B1g2 model for its similar number of parameters to ResNet50, which is the backbone of the state-of-the-art (SOTA) methods we compared with in Table 2. After testing multiple configurations, we found that a single final layer with 6 output neurons performed the best. The network was trained for 30 epochs using the Adam optimizer with initial learning rates of 1e-5 for the backbone and 1e-4 for the final fully connected layer, with both rates halved every 10 epochs. A batch size of 64 was used.

Our experimental evaluation involved using the synthetic 300W-LP dataset for training and AFLW2000 and BIWI real-world datasets for testing. Mean Absolute Error (MAE) of Euler angles was used as the evaluation metric, and our rotation matrix predictions were converted into Euler angles. As one can see in Table 2, our method outperformed the current state-of-the-art landmark-free approaches for head pose estimation, achieving almost 20% lower error rate on the AFLW2000 test dataset and the lowest error rate on all three rotation angles (yaw, pitch, and roll). On the BIWI dataset, our method achieved state-of-the-art results for the MAE. Our approach reported very balanced errors, indicating the network was able to learn consistently and robustly compared to other methods with diverging results on single angle errors.

5.2 Results for gaze estimation

The proposed L2CS-Net for gaze estimation was trained on the MPIIGaze datasets with Adam optimizer and a learning rate of 0.00001 for 50 epochs. We normalize the dataset images as in [24] to remove the effect of head pose

Table 4 Evaluation of our emotion prediction approach on AffectNet

	Approach	Class		valence		arousal	
		Acc	F_1	CCC	RMSE	CCC	RMSE
[46] ¹	SVM ¹	60	0.37	–	–	–	–
	MSCognitive ¹	68	0.51	–	–	–	–
	AlexNet ^{1,4}	64–72	0.55–0.57	–	–	–	–
	SVR ¹	–	–	0.340	0.494	0.199	0.400
	AlexNet ¹	–	–	0.541	0.394	0.450	0.402
Proposed multitask ²	MobileNet V3 (large)	70.550	0.447	0.735	0.338	0.413	0.261
	Mobileone (S0)	73.279	0.470	0.765	0.316	0.479	0.248
	Resnet18	71.606	0.477	0.751	0.326	0.420	0.258
	Resnet50d	74.048	0.502	0.770	0.313	0.479	0.249
	EfficientNet V2 (small)	73.946	0.518	0.793	0.304	0.489	0.250
	ConvNext (pico)	73.019	0.472	0.776	0.309	0.485	0.252
	Swin Transformer ³ (tiny)	75.654	0.514	0.753	0.310	0.467	0.244
	Swin Transformer ³ (base)	76.813	0.535	0.782	0.296	0.513	0.238

Bold values indicate the best result

¹ [46], trained on full training set and tested on unpublished test set

² Trained on 70% of the AffectNet training samples and tested on the remaining (test set still not published)

³ Patch=4, window=7

⁴ For CNN-based classification, Mollahosseini proposed 4 models with AlexNet backbone

variations. Table 3 shows the comparison of mean angular error between our proposed network and state-of-the-art methods on MPIIFaceGaze dataset. Our proposed network achieves state-of-the-art gaze performance with an angular error of 3.86 on MPIIFaceGaze.

5.3 Results for emotion prediction

Since the official test set of AffectNet is still not published, we split the available data randomly into 70% for training and keep the remaining for testing, similar to how Mollahosseini et al. split the data to the original train and test set [46]. Since AffectNet is relatively comprehensive, the results on our test set should be comparable, but one should note that they may suffer from reduced amount of training samples. For training, we use the following augmentation techniques: random crop, random erasing, color fitter, random grayscale and random horizontal flip. Then, the images are normalized to 224×224 pixels.

5.3.1 Evaluation of backbones for the multitask network

The proposed network was trained with the PyTorch framework using Adam optimizer with learning rates from 0.00001 to 0.001 and batch sizes from 8 to 64 depending on the backbone size and was trained for 50 epochs. We use the small validation set of AffectNet to estimate at which

epoch the model starts overfitting. As backbone, we compared various different established but also new models all pretrained on ImageNet (1k). Results are shown in Table 4, where we also compare with the baseline of Mollahosseini et al. who tested different single task models including SVM and SVR. We achieved best results with a vision transformer (base Swin Transformer), which, however, is also the biggest model that we have tested.

Considering the importance for our future works of fast prediction even on weak hardware, also the results for MobileNet and especially the new Mobileone architecture are promising. Mobileone is designed to mitigate the architectural and optimization bottlenecks present in other efficient neural networks, making it a highly effective backbone for efficient neural networks on mobile devices [14]. As one see in Table 4, the smallest version of this CNN (Mobileone S0, which has fewer parameter than all other tested models) outperforms MobileNet V3 and ResNet18 and almost reaches the performance of ResNet50d. However, MobileOne is designed for current-generation iPhones and runs much slower on a regular CPU (such as an Intel i7) than MobileNet.

Another interesting architecture is EfficientNet. EfficientNet is designed to be highly efficient in terms of both computational resources and model size. It achieves state-of-the-art performance on several image classification

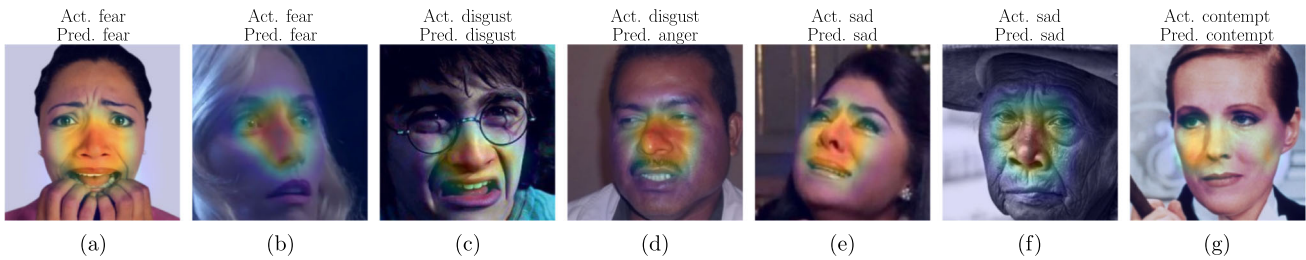


Fig. 8 Heatmaps show the areas of most relevant features for the predicted class for some samples of the AffectNet dataset

Prediction	neutral	14944 20.16	2022 2.73	2265 3.06	1246 1.68	329 0.44	170 0.23	1793 2.42	235 0.32	23004 31.95 35.84
	happy	2492 3.36	31641 42.69	262 0.35	674 0.91	83 0.11	111 0.15	330 0.45	450 0.61	36043 47.79 52.21
sad	682 0.92	114 0.15	3478 4.69	88 0.12	189 0.26	85 0.11	334 0.45	11 0.01	4981 69.83 80.17	
surprise	258 0.35	152 0.21	72 0.10	1229 1.66	303 0.41	15 0.02	56 0.08	5 0.01	2090 58.80 64.20	
fear	31 0.04	22 0.03	67 0.09	136 0.18	542 0.73	23 0.03	67 0.09		888 61.04 68.96	
disgust	53 0.07	64 0.09	61 0.08	15 0.02	32 0.04	332 0.45	158 0.21	6 0.01	721 46.05 53.95	
anger	891 1.20	113 0.15	418 0.56	118 0.16	135 0.18	207 0.28	3486 4.70	34 0.05	5402 64.53 70.47	
contempt	227 0.31	393 0.53	33 0.04	14 0.02	4 0.01	13 0.02	87 0.12	211 0.28	982 21.49 28.51	
sum_col	19578 76.33 81.67	34521 95.69 8.41	6656 52.25 17.75	3520 34.91 63.09	1617 33.52 66.48	956 34.73 65.27	6311 55.24 41.75	952 22.16 77.84	74111 75.39 23.62	
		neutral	happy	sad	surprise	fear	disgust	anger	contempt	sum_lin
		Actual								

Fig. 9 Confusion Matrix for the basic emotions (using a Vision-Transformer)

benchmarks such as ImageNet with significantly fewer parameters and floating-point operations (FLOPs) compared to other state-of-the-art models like ResNet50 [77]. When it comes to emotion recognition, this particular method demonstrates exceptional performance in

predicting valence, as indicated by its high CCC score. This is particularly relevant because valence provides insight into the degree to which a person is experiencing negative, positive, or neutral emotions. This is why we use EfficientNet to predict basic emotions and valence arousal values as features for deception detection.

Emotion Class Confusion In Fig. 9, the confusion matrix shows detailed classification results of the base Swin Transformer—the tested model with the best overall performance—on the AffectNet test set. Although contempt may be more meaningful than the other classes in the context of deception detection, its high confusion rate especially with the happy class renders it a less reliable feature for this purpose. Possible explanations for the high confusion between contempt and neutral or happy emotions are that the facial expressions for contempt can be quite similar to happy, and there are only a limited number of samples available for the contempt class.

Apart from the problems regarding contempt, the confusion matrix reveals another issue. AffectNet includes numerous instances where annotators had disagreements regarding the appropriate class label, potentially resulting in misclassification. This is particularly evident in cases where the facial expressions are subtle, making it difficult to differentiate between neutral and emotional states. On the other hand, such subtle facial expressions are important for real-life scenarios, where emotion can be mixed or of

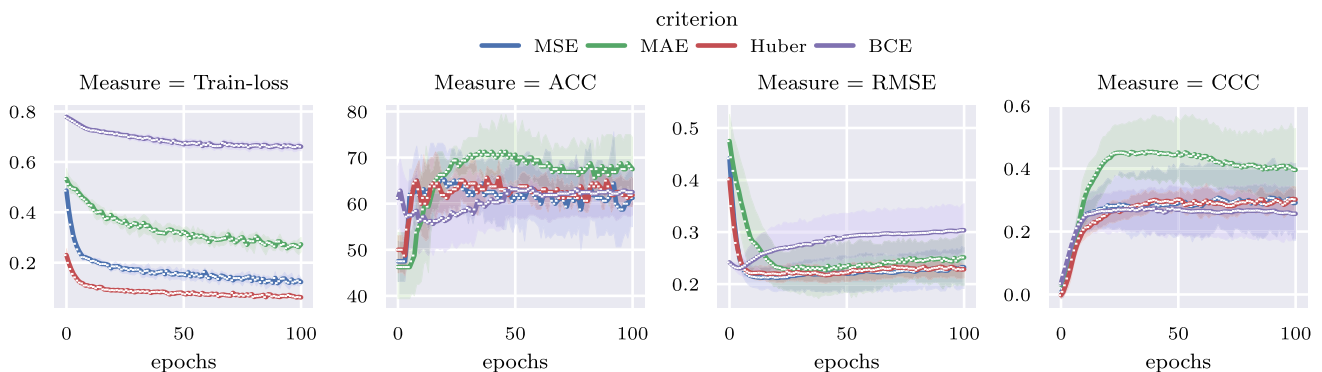


Fig. 10 Results of the LSTM training on the RL database

Table 5 Test results (average and standard error) using different optimizers for training

Optimizer	Train-loss	Accuracy	RMSE	CCC
Adam	.37 ± .00	65.59 ± 1.76	.24 ± .01	.40 ± .03
Nadam	.38 ± .00	65.91 ± 2.10	.24 ± .01	.40 ± .04
Adadelata	.21 ± .01	68.41 ± 1.92	.25 ± .01	.38 ± .03
Adagrad	.29 ± .00	68.05 ± 1.81	.23 ± .01	.41 ± .03
SGD	.50 ± .00	50.00 ± .00	.36 ± .01	.10 ± .01

Used loss: MAE, learning rate=0.00003 (0.003 for Adagrad, rho=0.9 for Adadelata), used dataset: RL

weak intensity. This highlights the significance of incorporating VA values as features for deception detection, which capture the intensity of the displayed emotions, in order to minimize confusion with neutral expressions.

In Fig. 8, we present some samples from the AffectNet dataset, wherein heatmaps are utilized to highlight the most significant regions contributing to the predicted emotion class. These areas represent the regions from which the most relevant features were derived, leading to the highest score for a particular class.

It is evident that these regions are consistently located on the face, particularly in areas where one or more Action Units (AUs) are present.

Furthermore, we can observe two samples with ambiguous expressions. In Fig. 8g, the expression was correctly identified as contempt. However, in the absence of context information, contempt is primarily recognizable by a wry smile (notably, the heatmap is asymmetric for this sample), which introduces a significant potential for confusion with neutral or happy (not only in making predictions but also in establishing manual ground truth). Additionally, in Fig. 8d, the classification was anger,

whereas the ground truth label is disgust. It is our belief that the image itself may suggest multiple valid interpretations. For comparison, Fig. 8c clearly depicts an expression of disgust.

5.4 Evaluation of the four available deception detection datasets

In this section, we will conduct experimental evaluations on the given four databases using discriminative classifiers and LSTMs models. These evaluations will be performed using multimodal features extracted from the deep learning approaches described earlier. Furthermore, we will proceed to conduct additional experiments using the data from our rolling dice Experiment (RDE).

5.4.1 LSTMs

It is important to note that approaches like LSTMs may struggle to learn meaningful patterns if the available datasets lack comprehensiveness. However, given the possibility of a temporal pattern existing in at least one of the datasets, which could indicate deception, we train LSTMs using the most prevalent criteria. This allows us to evaluate whether LSTMs perform better or worse compared to discriminative models in terms of detection accuracy. For the RL dataset, Fig. 10 shows the training losses of the used criteria and the validation results for different measures.

It is evident that Mean Absolute Error (MAE) outperforms other loss functions such as Binary Cross-Entropy (BCE) in terms of the evaluation metrics Accuracy and CCC (Concordance Correlation Coefficient). Despite initial expectations favoring BCE as the superior loss function, the results indicate that MAE yields better performance according to these evaluation metrics.

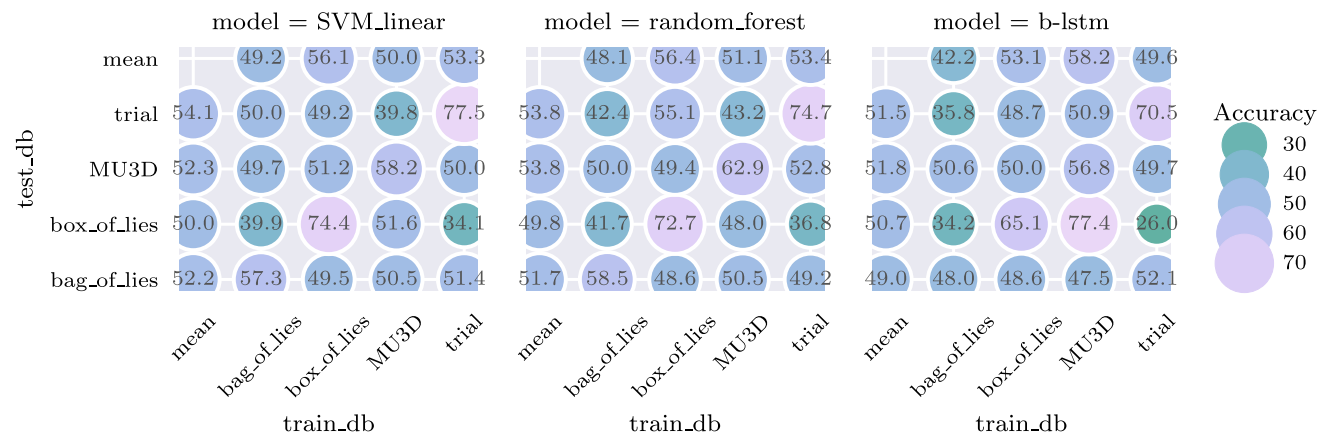


Fig. 11 Results for Intra- and Cross-database experiments for three different classifiers

MAE loss is computed as the average absolute difference between predicted and actual values. In the context of binary classification, the predicted values represent probabilities or scores indicating the likelihood of belonging to a specific class. One of the reasons why MAE performs well could be its robustness to outliers, unlike Mean Squared Error (MSE) or BCE loss functions. Considering the challenging nature of the problem and the limitations of the dataset, which may lead to an increased number of outliers, the resilience of MAE loss to these extreme values becomes advantageous. Furthermore, MAE loss equally considers both overestimations and underestimations, a crucial characteristic for accurate binary classification.

Overall, the performance of MAE over other measures suggests its suitability for the given binary classification task, highlighting its ability to handle outliers effectively and treat overestimations and underestimations equally.

We also explore the impact of different optimizers on the results. To determine the values of the measures, we select the epoch where the loss plateaus, suggesting the end of meaningful learning and the potential onset of overfitting. We then calculate the average of the loss values for the 10 preceding epochs. As illustrated in Table 5, the Adam optimizer, known for its effectiveness in numerous deep learning tasks, particularly for datasets exhibiting complex, nonlinear patterns, gives results which not significantly differ from those obtained using Adadelta or Adagrad. Conversely, SGD, acknowledged for its efficacy on large-scale datasets but prone to local minima entrapment, demonstrates less favorable outcomes.

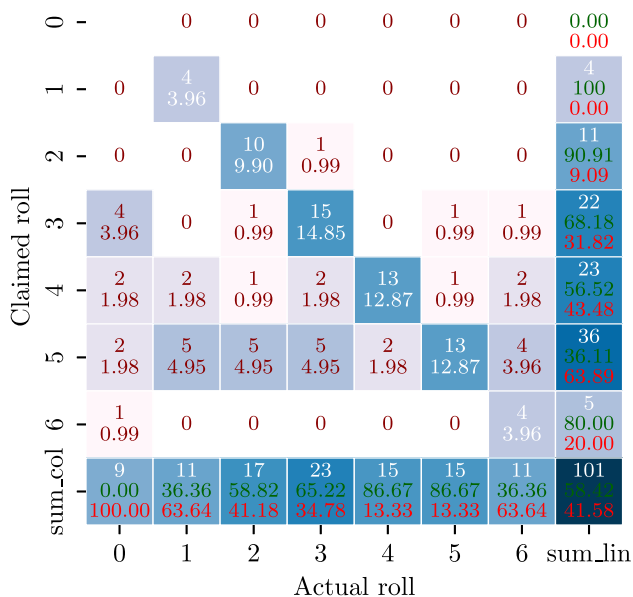


Fig. 12 The confusion matrix illustrates the subjects’ reported dice rolls after actually rolling a specific value between 1 and 6, or not rolling at all

In the upcoming section, we will demonstrate that discriminative models employing feature descriptors currently outperform LSTMs on the provided deception datasets, attributed to the limited training samples. Nevertheless, models like LSTMs offer enhanced flexibility for modification and scaling, possess the capability to independently learn temporal patterns (if available), and can be fine-tuned retrospectively with new data, such as scenarios where a baseline is accessible from test subjects. Furthermore, in pain recognition using a relatively larger database of approximately 20,000 samples, LSTM demonstrated slightly superior results compared to Random Forest (RF) using similar features (pose and AUs) [78].

5.4.2 Cross- and intra-database results

As mentioned earlier, the identification of deception is intricately tied to the specific problem being addressed. However, as far as our knowledge extends, no cross-database experiments have been conducted to assess this aspect qualitatively across the four publicly available deception databases.

Figure 11 shows the cross and intra-database results for SVM, RF and LSTMs. Cross-database results were obtained by training the model on all samples from one database and evaluating it on all samples from another. Conversely, intra-database results were obtained using cross-validation, specifically the repeated, stratified random train-test splits method with a total of 50 iterations for SVM and RF and 10 iterations for LSTMs, utilizing 70% of the samples for training purposes.

Cross-database Results As evident from the results in Fig. 11, most cross-database experiments yield poor outcomes, sometimes even worse than random guess.

In case of the BOX dataset, the underperformance can be attributed to the imbalanced distribution of truthful and deceptive samples when utilizing BOX as a training or test dataset. Moreover, the BOX database contains numerous surprisingly short true samples (1 or 2 s), necessitating a sequence length of ≤ 40 for LSTMs models, whereas the other datasets perform best with a length of 200. In fact, when employing LSTMs as a classifier in cross-database experiments, the excessively small samples from BOX must be excluded. This exclusion further exacerbates the dataset’s imbalance. This explains the relatively good results using MU3D for training, BOX for testing and LSTMs as classifier, where the model actually does not perform better than a trivial classifier.

Additionally, it is worth noting that the contribution of modalities varies across databases, which is another contributing factor to the unsatisfactory performance of cross-database results. This highlights the importance of training deception detection models on datasets that are specifically

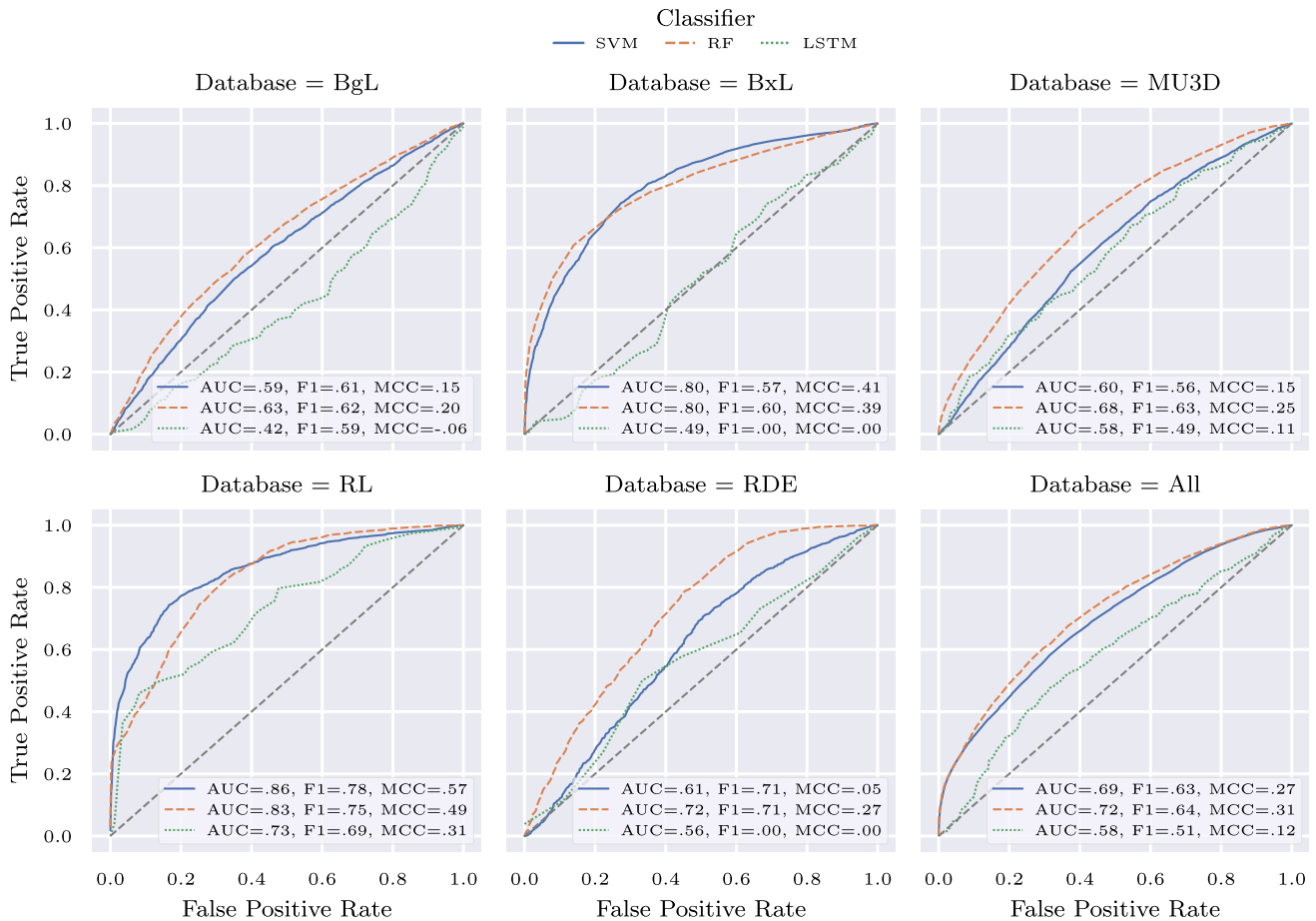


Fig. 13 Evaluation of the performance of classifiers using ROC curves, along with additional measures such as AUC and FI-Score, on various datasets. The ROC curves helped us assess the trade-off

between true positive rate and false positive rate at different classification thresholds. For this experiment, deception is used as positive class

tailored to the target scenario, emphasizing that the choice of the dataset is even more critical than in general machine learning problems.

Intra-database Results The intra-database results provide insight into the performance of the tested classifiers across the four datasets.

In general, the BOX dataset achieves the highest accuracy. However, it should be noted that this dataset is biased, which leads to classification results benefiting from a priori knowledge. For instance, a trivial classifier achieves an accuracy of 64% instead of the expected 50% for the RL and MU3D datasets and 52% for the BgL dataset. Specifically, the accuracy rates on BOX are 74% for SVM, 72% for RF, and 65% for LSTMs. However, it is crucial to consider the F1-scores, as they are less sensitive to biased class distributions. The F1-scores on BOX are 0.57, 0.60, and 0.0/NaN for SVM, RF and LSTMs classifiers, respectively. In the case of an F1-score of 0, it indicates that either the recall or sensitivity is 0, resulting in an

undefined value (NaN). This demonstrates that the LSTMs model did not learn meaningful feature patterns on this dataset, while the use of descriptors for feature extraction and classification using SVM and RF performs better. While LSTM exhibits the capacity to grasp intricate temporal patterns, its efficacy hinges on the availability of an extensive dataset. In domains characterized by highly individual challenges, such as deception detection, a larger sample size becomes imperative for the model to discern and generalize these nuanced patterns effectively, mitigating the risk of overfitting.

5.5 Evaluation of our rolling dice experiment

In addition to the cross-dataset analysis presented in the previous section, we further examined the performance of our proposed approach by investigating single modalities, alternative measures, and incorporating data from our RDE.

Table 6 Accuracy (and standard error) of each set of facial cues (features) for linear SVM and Random Forest

feature group		BgL	BxL	MU3D	RL	RDE
SVM ¹	gaze	54.63 ± 0.33	69.18 ± 0.37	52.84 ± 0.42	68.00 ± 0.68	58.06 ± 0.00
	AU	52.02 ± 0.15	65.88 ± 0.21	50.86 ± 0.35	64.56 ± 0.62	58.06 ± 0.00
	pose	51.96 ± 0.30	64.19 ± 0.01	53.10 ± 0.36	71.06 ± 0.62	58.03 ± 0.03
	emotion	53.32 ± 0.30	69.48 ± 0.28	52.93 ± 0.43	68.44 ± 0.63	59.74 ± 0.34
	all	57.16 ± 0.43	74.33 ± 0.41	58.25 ± 0.57	77.44 ± 0.59	62.90 ± 0.83
Random Forest ¹	all	55.18 ± 0.42	73.22 ± 0.42	63.17 ± 0.52	73.56 ± 0.57	67.96 ± 2.14
	gaze	55.66 ± 0.56	70.33 ± 0.60	57.00 ± 0.52	70.39 ± 0.82	54.77 ± 0.93
	AU	52.21 ± 0.59	71.88 ± 0.67	56.81 ± 0.59	66.67 ± 0.89	56.84 ± 0.92
	pose	56.15 ± 0.64	67.25 ± 0.63	55.42 ± 0.58	71.61 ± 0.86	59.68 ± 1.20
	emotion	53.85 ± 0.60	70.12 ± 0.58	61.81 ± 0.71	67.61 ± 0.92	69.48 ± 1.10
trivial classif.	51.60	64.20	50.00	50.00	58.06	
SOTA	49.9-55.26 ²	69 ³	—	60-75 ⁵	—	
manual	—	55 ³	52 ⁴	62 ⁵	—	

Highlighted are the best sets of cues separately for each database and classifier and—in the center of the table—the result for the best classifier for each database in the case of all feature groups. Performed statistical analysis: Paired *T* test passed: \curvearrowright ($p < 0.05$) \curvearrowleft , Wilcoxon test passed: \curvearrowright ($p < 0.05$) \curvearrowleft ($p < 0.1$), not passed: \curvearrowright \curvearrowleft .

¹Proposed approach

²Gupta et al. using video features [12] (EEG and eye tracking were tested on a subset only)

³Soldner et al. using manually extracted features plus RF [55]

⁴Lloyd et al. [56], 5Perez et al. [57]

First, let’s examine the RDE from an economic perspective. Figure 12 depicts the confusion between the actual and claimed values in our RDE, where participants were aware of being recorded and the records were subsequently analyzed by an external entity, separate from the experiment-conducting institution. The MTVE system was used for the experimental implementation, which was specifically designed to gather video data from communication-based experiments [79].

Notably, the percentage of subjects who were completely truthful increased to 58%, as compared to the 39% reported in the earlier study conducted by Fischbacher et al. without the presence of records [8]. This suggests that the documented record of deceptive behavior had a noticeable yet moderate impact on the participants’ behavior.

Furthermore, the findings indicate a strong correlation between the actual value of the dice roll and the participants’ behavior. Specifically, most individuals who rolled a 1 or 6 (representing the two worst possible outcomes) claimed they rolled a 4 or 5 (the two best possible outcomes). Conversely, most participants who rolled a 2, 3, 4, or 5 were truthful about their roll. In the end, out of the 101 subjects, 27 falsely claimed to have achieved a better roll

than they actually had, while 6 subjects claimed to have rolled worse than their actual result, and 59 called the actual value. Additionally, 9 participants did not roll at all (indicated with 0). These discoveries suggest that the accuracy of self-reported data in experiments can be influenced by various factors, including the presence of a record that exposes deception and the actual value of the outcome.

Please note that the recording also enables a more accurate calculation of the overall deception rate. If—like in [8]—no knowledge of the actual dice roll is available, the claimed roll can only be compared to a statistical distribution (which is an equal probability distribution in the case of a fair dice). However, using only a hundred or a few hundred samples, the actual distribution can clearly differ from the expected one, as it is the case in Fig. 12. Using 100 rolls, the average deviation from the ideal distribution is $18.24 \pm 5.2\%$, where \pm indicates the standard deviation for simulating the experiment 50 times. Fischbacher et al. conducted their main experiment with 389 rolls [8], resulting in a deviation of $8.6 \pm 3.1\%$. In our case, this deviation has no influence on the calculated deception rate, since we use the recorded actual rolls as ground truth.

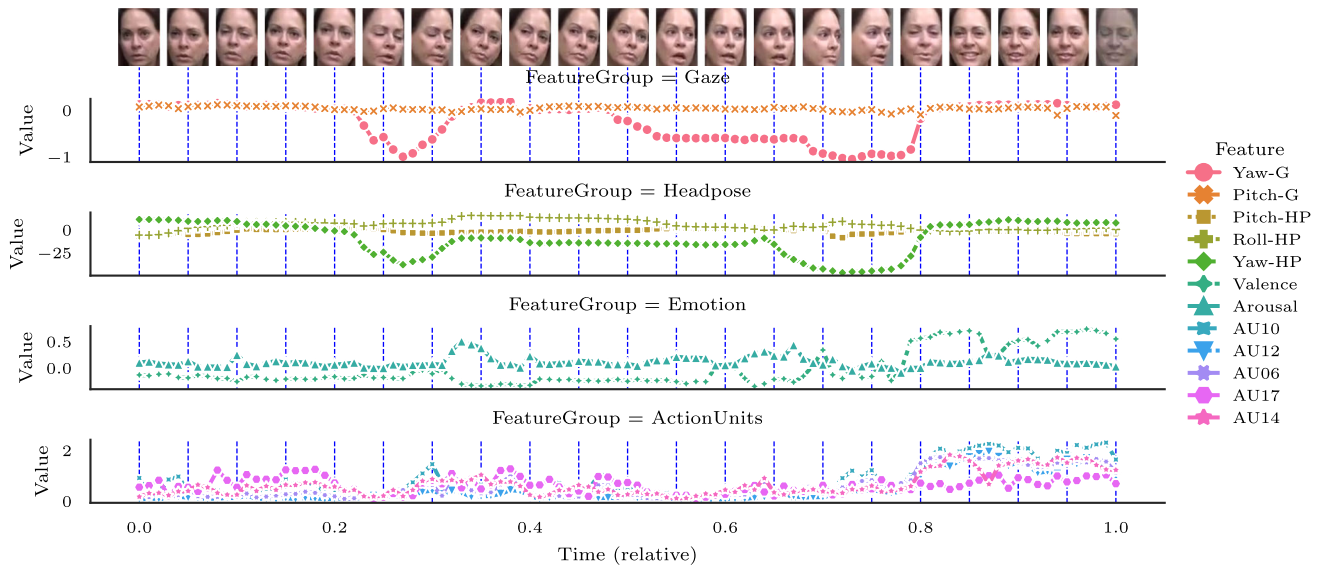


Fig. 14 Qualitative result shows the most relevant features over time for sample 47 of the lie set of the RL dataset (G=gaze in RAD, HP=head pose in °)

5.5.1 RoC curves and F1-scores

To evaluate our deception detection approach on the RDE and to compare the results with the four available datasets, we compute Receiver Operating Characteristic (RoC) curves, F1-scores and MCC measure for all three classifiers as shown in Fig. 13.

RoC curves are advantageous for evaluating model performance in binary classification problems with multiple datasets. They provide a visual representation of the trade-off between true positive and false positive rates, making them useful for comparing models across datasets with varying biases. The derived Area Under the Curve (AUC) measure allows comparing RoC curves of different models with a single scalar. On the other hand, F1-score combines precision and recall into a single metric, providing a balanced evaluation of the model's performance. MCC on the other hand takes into account true negatives and is more suitable for imbalanced datasets, as BOX and RDE. Utilizing RoC curves and F1-scores and MCC helps assess model effectiveness in the presence of biased and unbiased datasets.

It is evident that LSTMs perform worst over all in terms of AUC, F1-Score and MCC. In the sense of AUC, this is particularly true for BgL and BOX, while F1-Scores and MCC are 0 for the biased BOX and RDE. Although on BgL an F1 score of 0.59 is achieved, the AUC is less than 0.5 and the MCC is slightly negative, indicating that the classifier did not successfully learn meaningful features or that these features are not present in the test set. When

comparing the AUC and MCC scores of all four low-stake datasets, it is observed that the LSTMs model performs the best on the MU3D dataset. This result was expected since the MU3D dataset has the highest number of samples.

Results on the high-stake RL dataset are better, but still not good compared to the discriminative classifiers. SVM achieves the best results on RL for all measures, while RF exhibits a significantly higher AUC and MCC on RDE, indicating that RF exhibits a better ability to capture the overall performance of the classification task on this more challenging dataset. Moreover, when considering the results using data from all five datasets collectively, overall both discriminative classifiers yield comparable results, but RF may be slightly more accurate, as shown in Fig. 13.

Similar results were also reported in the literature. Monaro et al., in a study using unpublished data, compared various approaches [80]. They achieved an AUC of 0.72 ± 0.02 using Action Unit features from OpenFace and an SVM as a classifier. However, when the SVM was replaced with an LSTM, the AUC decreased to 0.57 ± 0.03 . Additionally, utilizing a 3D-CNN (C3D) resulted in an AUC of 0.64 ± 0.03 , still lower than the approach employing features and SVM. Rill et al. also employed a similar approach, as described in [81], which was tested on the RL dataset. They utilized all OpenFace features (Action Units, gaze, head pose, and facial landmarks) and achieved an AUC of 57.4 using SVM for classification, and 56.0 when employing an LSTM.

In conclusion, while LSTMs are often regarded as more sophisticated models compared to discriminative

classifiers, their limited suitability for our problem arises due to the insufficient amount of available data. Considering the challenging task of deception detection without control questions and contact-based sensors, results are promising, especially for the high stake RL dataset. In case of our RDE, detecting deception is much harder, as expected due to the low-stake context. However, compared to BgL and MU3D, results are good in sense of AUC, F1-score and MCC using RF as classifier.

5.5.2 Evaluation of single modalities

In this section, we explore the effectiveness of each set of facial cues—namely the outputs of the CNNs for gaze, AU, pose, and emotion—in distinguishing between deception and truth. Integrating every theoretically applicable modality into a single system, which includes specialized sensors like thermal data or contact-based modalities such as skin conductance or EEG, would have potential but also severe drawbacks. Such an approach may prove costly, impractical, challenging, or impossible to train on datasets without such highly specialized sensor data. Conversely, relying solely on a single modality/set, such as only Action Units, may not fully utilize the potential information provided by the used device, which typically includes head-centric RGB-video, as observed in all used deception datasets. Therefore, we will assess the performance of both single and combined sets across the five datasets utilized in our analysis.

Table 6 shows results for single and combined sets for all datasets using SVM and RF. (As shown above, LSTMs is not working well on the limited number of samples, so we skip it for this experiment.) To analyze the significance, we employed a pairwise *t* test and a Wilcoxon test, both visualized with connectors. The *t* test was utilized under the condition that both variables exhibited a Gaussian-like distribution, as confirmed by the Shapiro test, and their variances passed the Levene test. Conversely, in cases where these assumptions were not met, we used the Wilcoxon test instead, which is a nonparametric method robust against such key assumptions.

We pairwise compare the set of facial cues that are achieving the highest accuracy with the other sets. As evident from the results, the performance of the modalities varies depending on the dataset used. Notably, the sets of emotion cues and head pose achieved the highest accuracy four times each, whereas AU and gaze outperformed the other sets only once. Restricting our analysis to cases where all *t*/Wilcoxon tests passed, we observed that the set of emotion cues surpassed the other sets in three instances while pose, AU, and gaze achieved the highest accuracy

just once. Analyzing the contribution of the base features of the emotion cues, we identified fear, disgust, and arousal as particularly relevant.

Comparing the classifiers using all sets of facial cues, SVM outperforms RF on three databases. However, in one case Wilcoxon test is only passed with a *p* value of 0.1, suggesting an overall similar performance level for both classifiers. Notably, employing all sets of facial cues significantly outperformed using any single set with the same classifier. Only in 2 out of 10 cases did a single set demonstrate higher accuracy; however, these cases did not achieve statistical significance according to the *t* test.

Overall, it can be concluded that the set of emotion cues exhibits the highest potential on most deception datasets, but using combined sets makes the system more robust. However, it is reasonable to assume that the recognition rates could be further enhanced by incorporating additional modalities, such as those based on audio sensors. Non-contact vital parameter estimation, such as monitoring heart rate or changes in blood circulation over time, may also prove valuable, although it may encounter limitations in certain scenarios. These limitations include very short samples, variations in lighting conditions, and especially video compression, which can impact the functionality of such modalities.

5.5.3 Qualitative results

Finally, Fig. 14 presents a qualitative result from a single sample of the RL dataset. The figure displays the basic features of all modalities over time, along with snapshots of the video data taken at regular intervals. (Basic emotions are omitted for clarity.)

In this particular sample, where the subject exhibits deceptive behavior, there is an observed increase in valence during the last quarter of the video. This observation can be seen as an example of “liar’s delight,” wherein individuals derive a sense of satisfaction or enjoyment from engaging in deceptive behavior.

Additionally, there is a noticeable change in yaw of gaze and head pose, suggesting that the subject may be intentionally avoiding eye contact with the judge or lawyer. While the assumption concerning the subject’s motivation is challenging to verify definitively, our prior findings highlight the increased relevance of gaze and head pose within the RL dataset compared to other datasets. This observation lends support to the hypothesis that the subject’s avoidance of eye contact could be associated with deceptive behavior.

Please note that this qualitative result serves as an illustrative example, showcasing intriguing behavioral

patterns and features over time. It is important to acknowledge that not all samples exhibit such pronounced behavior, particularly in the case of low-stake datasets. Extracting indicative features for deception in average examples may present a greater challenge.

6 Conclusion

In this study, we proposed an approach for deception detection based on several sets of facial cues, with each set corresponding to a specific visual modality predicted by its own Convolutional Neural Network (CNN).

Initially, we developed approaches utilizing CNN models to compute intermediate results for the individual modalities of gaze, (head) pose, and emotion as a modality of facial expressions. Comprehensive datasets are available for all these modalities, and we compared our models with the state-of-the-art, achieving good results. Further, we used a pre-trained CNN for the recognition of Action Units (AUs) as a second modality of facial expressions. Subsequently, the CNN outputs were utilized for early fusion in deception classification. We trained and tested this approach on four established deception datasets and found that the context of such datasets is crucial, with the best performance achieved on the only high-stake dataset (despite its lower technical quality regarding resolution, noise, etc.).

Discriminative classifiers such as Support Vector Machine (SVM) or Random Forest (RF) were found to be more effective for the fusion step than Long Short-Term Memory networks (LSTMs) across all datasets, a finding supported by Monaro et al. on a different (unpublished) deception dataset [80]. The main reason is the severely limited sample size compared to other domains. Regarding the utilized modalities, we found that emotion—which encompasses the basic emotion class and regression values for valence and arousal—is overall best suited for concluding the truthfulness of a video statement. However, accuracy can be improved by incorporating multiple modalities in most cases.

Furthermore, we performed a new rolling dice experiment (RDE), generating an interdisciplinary dataset with a low-stake economic context. Previously, subjects were not recorded in the existing RDEs. We found that recording subjects mildly increases the percentage of honest subjects. More important, recording enables the evaluation of the proposed deception detection approach. We observed that deception detection in this scenario is challenging due to limited signs of guilt or other physical reactions. This observation applies similarly to other low-stakes datasets.

However, by employing multiple modalities and classifiers, we achieved an average accuracy of 67% and an F1-score of 0.71.

6.1 Limitations and future works

After all, we believe that the highly individual nature of physical reactions to deception, combined with the severe limitation of samples in existing datasets (less than 350), underscores the pressing need for a more comprehensive deception dataset. Such a dataset would enhance the outcomes of future studies by facilitating the training of more robust and generalized models for deception classification. We propose that such a dataset should encompass mid-stake scenarios, rather than solely focusing on extremely low-stake situations with minimal risk or motivation to lie. By incorporating scenarios that closely resemble real-world contexts, such as deception occurring in online conferences, we can more accurately capture the intricacies and subtleties of deceptive behavior.

To address this issue, we plan to acquire a dataset of at least double the size in the sense of samples compared to the currently most comprehensive dataset (Bag-of-lies and MU3D), where the proposed RDE was used as a pilot study. To ensure that the stake of the scenario is not unnecessarily low, we will design the experiment in a way where subjects have to lie to each other (which might cause more guilt than lying to a neutral observer), whereby the payout depends (apart from a small show up fee) on whether a “customer” subject can be convinced to buy a product. This way, we expect to achieve an increased stake compared to former datasets, where successful as well as failed attempts of deception have fewer positive, respectively, negative consequences. The dataset will be made freely available for research.

Regarding the proposed approach of deception detection, we have so far focussed on facial cues, which are known to be suitable for detecting patterns in emotional states (that are, for example, caused by guilt). Incorporating additional modalities such as contact-free vital parameters (heart and aspiration rate), audio features or hand gestures has the potential to enhance accuracy. It is important to note, however, that achieving very high accuracies, such as the claimed 90% accuracy in polygraph tests, is unlikely in fully automated low-stake scenarios that do not involve contact-based modalities, interrogation techniques, and context or person-specific knowledge.

Appendix A

Feature distribution and relevance

In this section, we present extended plots that provide an overview of multiple datasets and all features. While these plots may not be essential for understanding the main paper, they offer additional insights and details for interested readers. See Figs. 15, 16, 17

Fig. 16 Feature relevance computed by feature permutation using traditional classifiers (SVM and Random Forest). Features are ordered according to relevance for the trial database based on the SVM classifier. As one can see, the head pose and gaze features are especially relevant in case of the trial database, but the feature relevance is very sensitive to the dataset. Furthermore, the relevance also depends on the classifier

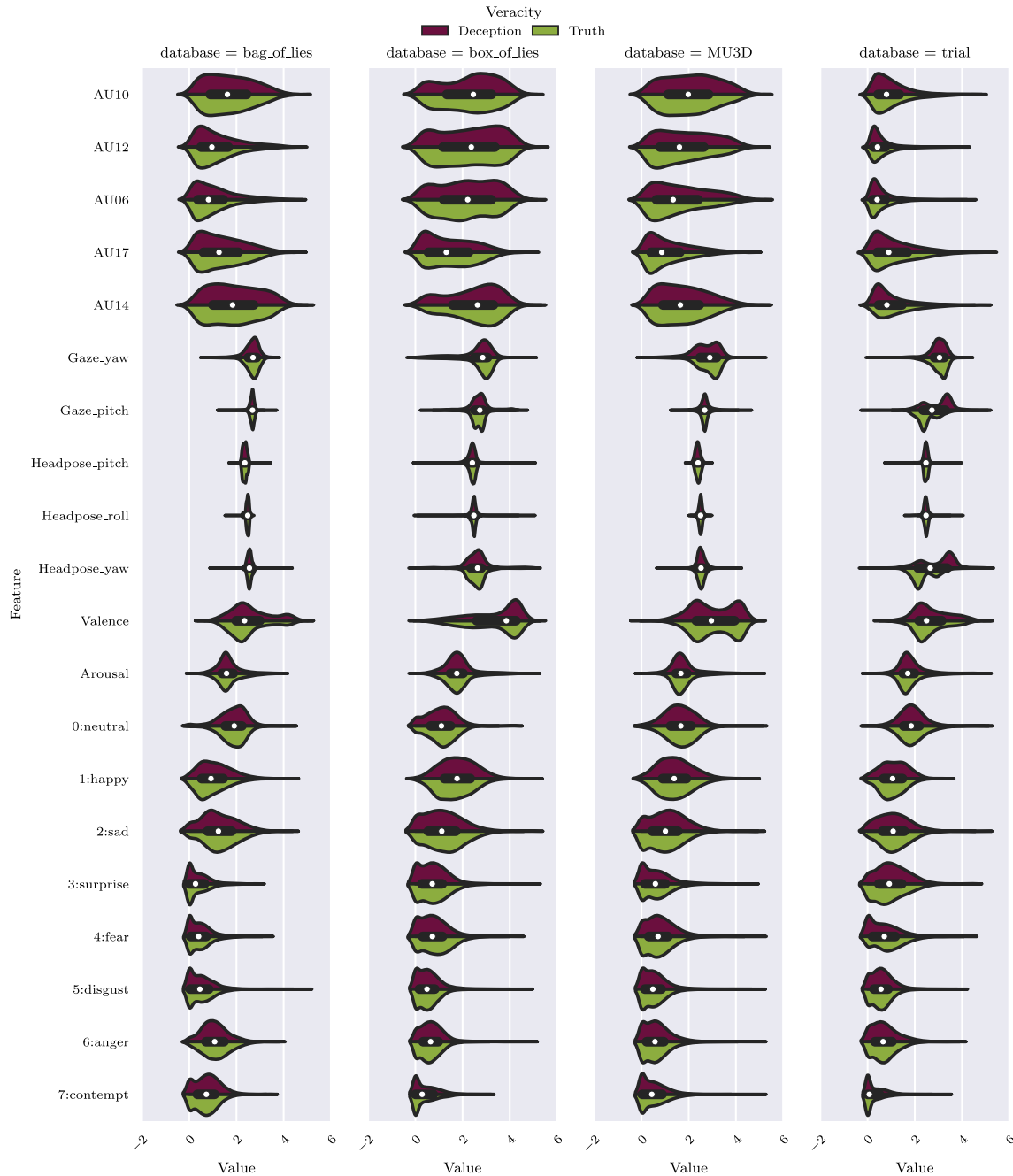
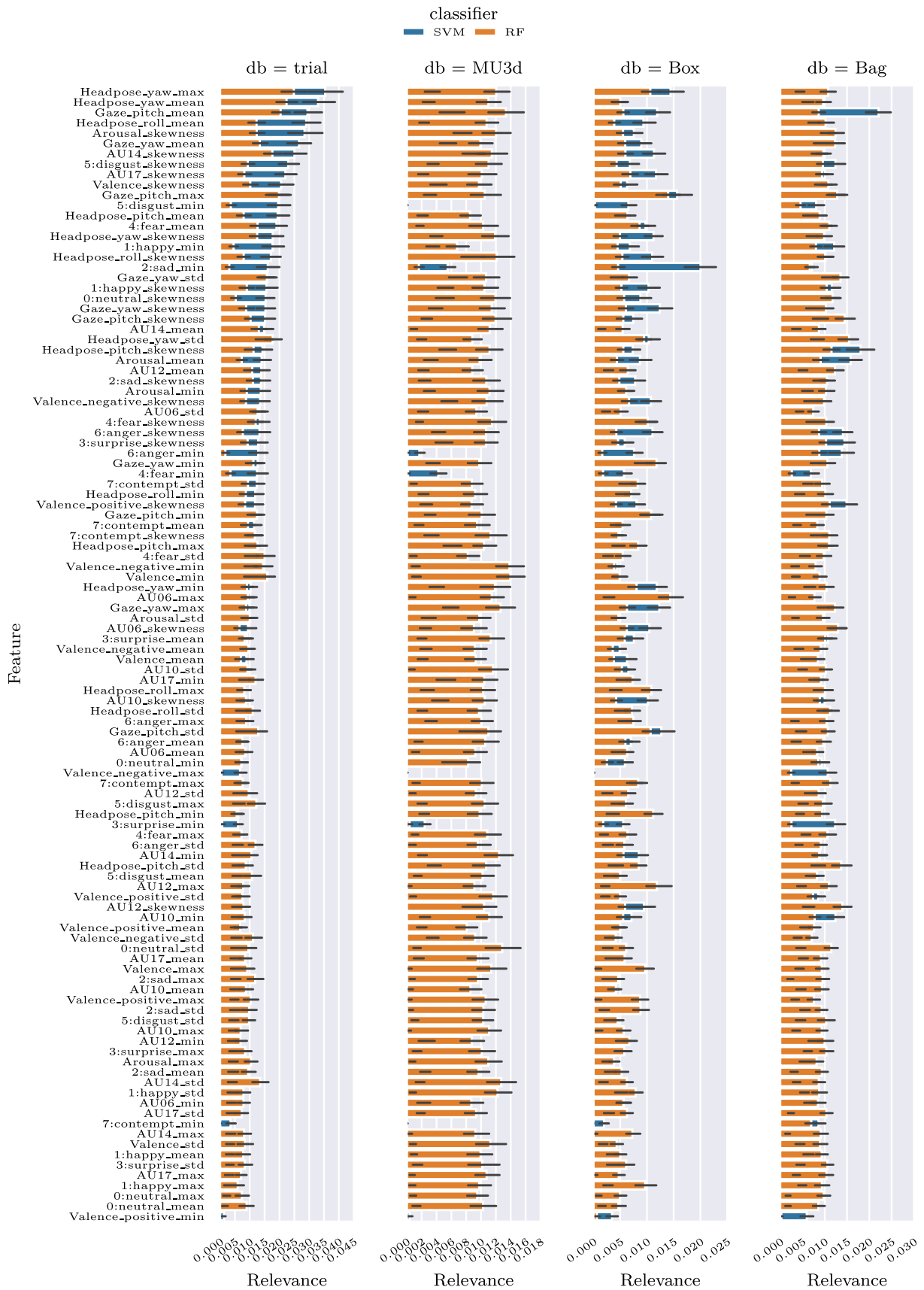


Fig. 15 Violin plots show the normalized distribution for all facial cues



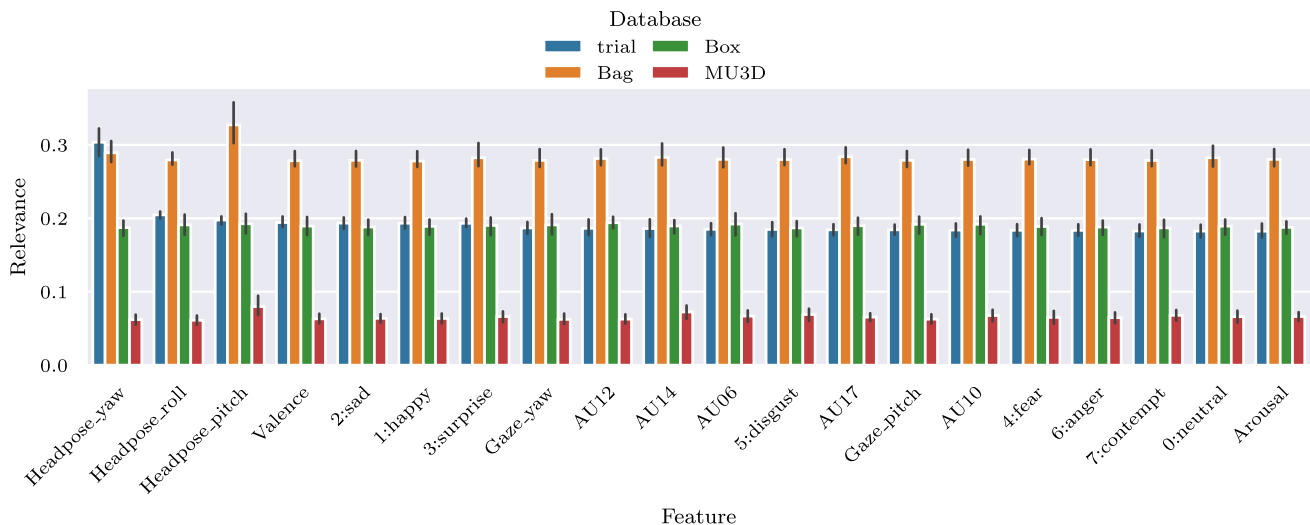


Fig. 17 Feature relevance for LSTM (ordered by Feature relevance for trial database, which is the only high-stake dataset). Feature relevance was computed by iteratively setting values for one feature to random values of its (Gaussian) distribution. As one can see, only

Author contributions Conceptualization was done by Laslo Dinges, and Marc Fiedler; Methodology was done by Laslo Dinges; Software was done by Laslo Dinges, Thorsten Hempel, and Ahmed Abdelrahmann; Dataset acquisition and labeling were done by Joachim Weimann, Dmitri Bershadskyy, Laslo Dinges, and Ayoub Al-Hamadi; Project heading was done by Joachim Weimann and Ayoub Al-Hamadi; Funding acquisition was done by Ayoub Al-Hamadi; Writing—original draft preparation was done by Laslo Dinges; Writing—review and editing was done by Laslo Dinges, Marc Fiedler, Dmitri Bershadskyy, and Johann Steiner; Supervision was done by Ayoub Al-Hamadi.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was funded by the German Research Foundation (DFG) project AL 638/13-1, AL 638/14-1 and AL 638/15-1 and by the Federal Ministry of Education and Research of Germany (BMBF) project AutoKoWaT, no. 13N16336. Open Access funding enabled and organized by Projekt DEAL. We would like to extend our sincere gratitude to Ms. Nina Ostermaier, who conducted the 'rolling dice' experiments as a valued member of our collaborative project team.

Data availability statement The publicly available datasets used for this research are described in [12, 55, 56] and [57]. The RDE dataset cannot be published due to data protection reasons.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

Ethical approval Concerning the rolling dice Experiment: Ethical review by the Society for Experimental Economic Research (GfeW). The GfeW with identification code (fnax4PA2) certified this experiment.

Consent of participant We obtained informed consent for all participants.

the head pose (yaw) feature has a clearly higher relevance than the other features in the case of the trial database, while features have a similar relevance in all other cases

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Alaskar H, Sbaï Z, Khan W, Hussain A, Alrawais A (2023) Intelligent techniques for deception detection: a survey and critical study. *Soft Comput* 27(7):3581–3600
- Sánchez-Monedero J, Dencik L (2022) The politics of deceptive borders: 'biomarkers of deceit' and the case of iborderctrl. *Inf Commun Soc* 25(3):413–430
- Brynjolfsson E, Wang C, Zhang X (2021) The economics of it and digitization: eight questions for research. *MIS Q* 45(1):473–477
- Weimann J, Brosig-Koch J et al (2019) *Methods in experimental economics*. Springer, Berlin
- Camerer CF, Dreber A, Johannesson M (2019) Replication and other practices for improving scientific quality in experimental economics. *Handbook of research methods and applications in experimental economics*. Edward Elgar Publishing, Cheltenham, pp 83–102
- Butski NR, Boostrom JRE, Bonds R (2022) An honest discussion about lies: how components of a seller's lies can impact customer relationship commitment. *Market Manage J* 32(1):17
- Taylor JB, Beach SR, Kontos N (2017) The therapeutic discharge: an approach to dealing with deceptive patients. *Gen Hosp Psychiatry* 46:74–78

8. Fischbacher U, Föllmi-Heusi F (2013) Lies in disguise—an experimental study on cheating. *J Eur Econ Assoc* 11(3):525–547
9. Nortje A, Tredoux C (2019) How good are we at detecting deception? A review of current techniques and theories. *South Afr J Psychol* 49(4):491–504
10. Herbig F (2020) The ‘psychology’ of polygraph’: engendering differential salience-concerns and caveats. *J Psychol Res* 2(2):1–10
11. Zloteanu M (2020) Reconsidering facial expressions and deception detection. *Handbook Facial Expr Emotion* 3:238–284
12. Gupta V, Agarwal M, Arora M, Chakraborty T, Singh R, Vatsa M (2019) Bag-of-lies: a multimodal dataset for deception detection. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp 0–0
13. Mattes K, Popova V, Evans JR (2023) Deception detection in politics: can voters tell when politicians are lying? *Polit Behav* 45(1):395–418
14. Vasu PKA, Gabriel J, Zhu J, Tuzel O, Ranjan A (2022) An improved one millisecond mobile backbone. *arXiv preprint arXiv:2206.04040*.
15. Junger M, Koning L, Hartel P, Veldkamp B (2023) In their own words: deception detection by victims and near victims of fraud. *Front Psychol* 14:1135369
16. Saini R, Rani P (2022) Ldm: a systematic review on lie detection methodologies.
17. Avola D, Cinque L, De Marsico M, Fagioli A, Foresti GL (2020) Lietome: preliminary study on hand gestures for deception detection via fisher-LSTM. *Pattern Recogn Lett* 138:455–461
18. Schepisi M, Porciello G, Aglioti SM, Panasiti MS (2020) Oculomotor behavior tracks the effect of ideological priming on deception. *Sci Rep* 10(1):9555
19. Werner P, Al-Hamadi A, Niese R, Walter S, Gruss S, Traue HC (2013) “Towards pain monitoring: Facial expression, head pose, a new database, an automatic system and remaining challenges. In: *Proceedings of the British Machine Vision Conference*. Citeseer, pp 1–13
20. Werner P, Saxen F, Al-Hamadi A (2017) Landmark based head pose estimation benchmark and method. *IEEE Int Conf Image Process (ICIP)* 2017:3909–3913
21. Ruiz N, Chong E, Rehg JM (2018) Fine-grained head pose estimation without keypoints. In: *2018 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 2155–215 509.
22. Hsu H-W, Wu T-Y, Wan S, Wong WH, Lee C-Y (2019) Quatnet: quaternion-based head pose estimation with multiregression loss. *IEEE Trans Multimed* 21(4):1035–1046
23. Huang B, Chen R, Xu W, Zhou Q (2020) Improving head pose estimation using two-stage ensembles with top-k regression. *Image Vis Comput* 93:103827
24. Zhang X, Sugano Y, Fritz M, Bulling A (2017) Mpiigaze: real-world dataset and deep appearance-based gaze estimation. *IEEE Trans Pattern Anal Mach Intell* 41(1):162–175
25. Xiong Y, Kim HJ, Singh V (2019) Mixed effects neural networks (menets) with applications to gaze estimation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7743–7752
26. Kellnhofer P, Recasens A, Stent S, Matusik W, Torralba A (2019) Gaze360: physically unconstrained gaze estimation in the wild. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6912–6921
27. Biswas et al P (2021) Appearance-based gaze estimation using attention and difference mechanism. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3143–3152
28. Kumar S, Bai C, Subrahmanian V, Leskovec J (2021) Deception detection in group video conversations using dynamic interaction networks. In: *Proceedings of the international AAAI conference on web and social media*, vol 15, pp 339–350
29. Pasquali D, Gonzalez-Billandon J, Aroyo AM, Sandini G, Sciutti A, Rea F (2021) Detecting lies in a child (robot)’s play: gaze-based lie detection in HRI. *Int J Soc Robot* 15:1–16
30. Constâncio AS, Tsunoda DF, Silva HdFN, Silveira JMd, Carvalho DR (2023) Deception detection with machine learning: a systematic review and statistical analysis. *Plos one* 18(2):e0281323
31. Ben X, Ren Y, Zhang J, Wang S-J, Kpalma K, Meng W, Liu Y-J (2021) Video-based facial micro-expression analysis: a survey of datasets, features and algorithms. *IEEE Trans Pattern Anal Mach Intell* 44(9):5826–5846
32. Wu Z, Singh B, Davis L, Subrahmanian V (2018) Deception detection in videos. In: *Proceedings of the AAAI conference on artificial intelligence*, vol 32
33. Frank MG, Svetieva E (2015) Microexpressions and deception. *Understanding facial expressions in communication: cross-cultural and multidisciplinary perspectives*, pp 227–242
34. Choi DY, Kim DH, Song BC (2018) Recognizing fine facial micro-expressions using two-dimensional landmark feature. In: *2018 25th IEEE international conference on image processing (ICIP)*. IEEE, pp 1962–1966
35. Jordan S, Brimbal L, Wallace DB, Kassin SM, Hartwig M, Street CN (2019) A test of the micro-expressions training tool: does it improve lie detection? *J Investig Psychol Offend Profil* 16(3):222–235
36. Talluri KK, Fiedler M-A, Al-Hamadi A (2022) Deep 3d convolutional neural network for facial micro-expression analysis from video images. *Appl Sci* 12(21):11078
37. Zhang L, Arandjelović O (2021) Review of automatic microexpression recognition in the past decade. *Machine Learn Knowl Extract* 3(2):414–434
38. Zhi R, Liu M, Zhang D (2020) A comprehensive survey on automatic facial action unit analysis. *Vis Comput* 36:1067–1093
39. Russell JA (1980) A circumplex model of affect. *J Pers Soc Psychol* 39(6):1161
40. Werner P, Handrich S, Al-Hamadi A (2017) Facial action unit intensity estimation and feature relevance visualization with random regression forests. In: *2017 seventh international conference on affective computing and intelligent interaction (ACII)*. IEEE, pp 401–406
41. Vinkemeier D, Valstar, J, Gratch (2018) Predicting folds in poker using action unit detectors and decision trees. In: *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, pp 504–511
42. P. Werner, F. Saxen, and A. Al-Hamadi (2020) Facial action unit recognition in the wild with multi-task CNN self-training for the emotionnet challenge. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp 410–411
43. Handrich S, Dinges L, Al-Hamadi A, Werner P, Al Aghbari Z (2020) Simultaneous prediction of valence/arousal and emotions on affectnet, aff-wild and afew-va. In: *Procedia computer science*, vol 170, pp 634–641
44. Chang WY, Hsu SH, Chien JH (2017) Fatauva-net: an integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp 17–25
45. Khorrami P, Paine T, Huang T (2015) Do deep neural networks learn facial action units when doing expression recognition?. In: *Proceedings of the IEEE international conference on computer vision workshops*, pp 19–27

46. Mollahosseini A, Hasani B, Mahoor MH (2017) Affectnet: a database for facial expression, valence, and arousal computing in the wild. *IEEE Trans Affect Comput* 10(1):18–31
47. Zhang L, Peng S, Winkler S (2019) Persemon: a deep network for joint analysis of apparent personality, emotion and their relationship. *IEEE Trans Affect Comput* 13(1):298–305
48. Li J, Chen Y, Xiao S, Zhao J, Roy S, Feng J, Yan S, Sim T (2017) Estimation of affective level in the wild with multiple memory networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp 1–8
49. Chu WS, De la Torre F, Cohn JF (2017) Learning spatial and temporal cues for multi-label facial action unit detection. In: *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*. IEEE, , pp 25–32
50. Avola D, Cinque L, De Marsico M, Di Mambro A, A. Fagioli, G. L. Foresti, R. Lanzino, and F. Scarcello (2023) Lietome: an lstm-based method for deception detection by hand movements. In: *International conference on image analysis and processing*. Springer, pp 387–398
51. S. Satpathi, K. M. I. Y. Arafath, A. Routray, and P. S. Satpathi (2020) Detection of deceit from thermal videos on real crime database. In: *2020 11th international conference on computing, communication and networking technologies (ICCCNT)*. IEEE, pp 1–6
52. Fiedler M-A, Rapczyński M, Al-Hamadi A (2021) Facial video-based respiratory rate recognition interpolating pulsatile ppg rise and fall times. In: *IEEE 18th international symposium on biomedical imaging (ISBI)*. IEEE 2021, pp 545–549
53. Ioannidis K, Offerman T, Sloof R (2022) Lie detection: a strategic analysis of the verifiability approach. *Am Law Econ Rev* 24(2):659–705
54. Bond GD, Lee AY (2005) Language of lies in prison: linguistic classification of prisoners' truthful and deceptive natural language. *Appl Cogn Psychol* 19(3):313–329
55. F. Soldner, V. Pérez-Rosas, and R. Mihalcea (2019) Box of lies: multimodal deception detection in dialogues. In: *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, Vol 1 (Long and Short Papers)*, pp 1768–1777
56. Lloyd EP, Deska JC, Hugenberg K, McConnell AR, Humphrey BT, Kunstman JW (2019) Miami university deception detection database. *Behav Res Methods* 51:429–439
57. V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo (2015) Deception detection using real-life trial data. In: *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pp 59–66
58. T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi (2022) 6d rotation representation for unconstrained head pose estimation. In: *IEEE international conference on image processing (ICIP)*. IEEE, pp 2496–2500
59. A. A. Abdelrahman, T. Hempel, A. Khalifa, and A. Al-Hamadi (2022) L2cs-net: Fine-grained gaze estimation in unconstrained environments. *arXiv preprint arXiv:2203.03339*
60. Y. Fan, J. Lam, and V. Li (2020) Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution. In: *Thirty-Fourth AAAI conference on artificial intelligence*
61. C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez (2016) Emotionet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 5562–5570
62. J. Deng, J. Guo, Z. Yuxiang, J. Yu, I. Kotsia, and S. Zafeiriou (2019) Retinaface: single-stage dense face localisation in the wild. In: *arxiv*
63. T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, and Y.-Y. Chuang (2019) Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*
64. Y. Zhou and J. Gregson (2020) Whenet: real-time fine-grained estimation for wide range head pose. In: *31st British machine vision conference 2020, BMVC 2020, virtual Event, UK, September 7-10. BMVA Press, 2020*. [Online]. Available: <https://www.bmvc2020-conference.com/assets/papers/0907.pdf>
65. Z. Cao, Z. Chu, D. Liu, and Y. Chen (2021) A vector-based representation to enhance head pose estimation. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)*, pp 1188–1197
66. H. Zhang, M. Wang, Y. Liu, and Y. Yuan (2020) Fdn: Feature decoupling network for head pose estimation. In: *AAAI*
67. K. Krafska, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba (2016) Eye tracking for everyone. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 2176–2184
68. Y. Xiong, H. J. Kim, and V. Singh (2019) Mixed effects neural networks (menets) with applications to gaze estimation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 7743–7752
69. X. Zhang, Y. Sugano, M. Fritz, and A. Bulling (2017) It's written all over your face: Full-face appearance-based gaze estimation. In: *Computer vision and pattern recognition workshops (CVPRW), 2017 IEEE conference on*. IEEE, pp 2299–2308
70. Z. Chen and B. E. Shi (2018) Appearance-based gaze estimation using dilated-convolutions. In: *Asian conference on computer vision*. Springer, pp 309–324
71. T. Fischer, H. J. Chang, and Y. Demiris (2018) Rt-gene: Real-time eye gaze estimation in natural environments. In: *Proceedings of the European conference on computer vision (ECCV)*, pp 334–352
72. Z. Chen and B. E. Shi (2020) Geddnet: a network for gaze estimation with dilation and decomposition. *arXiv preprint arXiv:2001.09284*
73. K. Wang, R. Zhao, H. Su, and Q. Ji (2019) Generalizing eye tracking with bayesian adversarial learning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 11 907–11 916
74. Cheng Y, Zhang X, Lu F, Sato Y (2020) Gaze estimation by exploring two-eye asymmetry. *IEEE Trans Image Process* 29:5259–5272
75. Y. Cheng, S. Huang, F. Wang, C. Qian, and F. Lu (2020) A coarse-to-fine adaptive network for appearance-based gaze estimation. In: *Proceedings of the AAAI conference on artificial intelligence, vol 34, no. 07*, pp 10 623–10 630
76. P. Biswas et al., (2021) Appearance-based gaze estimation using attention and difference mechanism. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 3143–3152
77. B. Koonce and B. Koonce (2021) Efficientnet, convolutional neural networks with swift for tensorflow: image recognition and dataset categorization, pp. 109–123
78. Othman E, Werner P, Saxen F, Al-Hamadi A, Gruss S, Walter S (2023) Classification networks for continuous automatic pain intensity monitoring in video using facial expression on the x-ite pain database. *J Vis Commun Image Represent* 91:103743
79. Bershadskyy D, Ghadwal S, Greif J (2022) MTV-Magdeburg tool for videoconferences. *Working Paper Series*
80. Monaro M, Maldera S, Scarpazza C, Sartori G, Navarin N (2022) Detecting deception through facial expressions in a dataset of videotaped interviews: a comparison between human judges and machine learning models. *Comput Hum Behav* 127:107063

81. R. Rill-García, H. Jair Escalante, L. Villasenor-Pineda, and V. Reyes-Meza (2019) High-level features for multimodal deception detection in videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 0–0

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Laslo Dinges¹  · Marc-André Fiedler¹ · Ayoub Al-Hamadi¹ · Thorsten Hempel¹ · Ahmed Abdelrahman¹ · Joachim Weimann² · Dmitri Bershadskey² · Johann Steiner³

✉ Laslo Dinges
laslo.dinges@ovgu.de

Marc-André Fiedler
Marc-Andre.Fiedler@ovgu.de

Ayoub Al-Hamadi
Ayoub.Al-Hamadi@ovgu.de

Thorsten Hempel
Thorsten.Hempel@ovgu.de

Ahmed Abdelrahman
Ahmed.Abdelrahman@ovgu.de

Joachim Weimann
Joachim.Weimann@ovgu.de

Dmitri Bershadskey
Dmitri.Bershadskey@ovgu.de

Johann Steiner
Johann.Steiner@ovgu.de

¹ Neuro-Information Technology Group, Otto-von-Guericke University, P.O. Box 4210, D-39106 Magdeburg, Saxony-Anhalt, Germany

² Faculty of Economics and Management, Otto-von-Guericke University, P.O. Box 4210, D-39106 Magdeburg, Saxony-Anhalt, Germany

³ Department of Psychiatry and Psychotherapy, University Hospital Magdeburg, Leipziger Str. 44, D-39120 Magdeburg, Saxony-Anhalt, Germany