



A Comparative Study on Face/Body-Based Lie Detection Using Convolutional Neural Networks

El-Sayed M. El-Alfy ^{1,2,3,*}, Abdelrahman S. Shbair¹, Omar A. Al-Oumi¹, Abdullatif M. Alsaad¹,
and Sadam Al-Azani ⁴

¹ Information and Computer Science Department,
King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

² Computer Engineering Department,
King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

³ IRC of Intelligent Secure Systems, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

⁴ SDAIA-KFUPM JRC for Artificial Intelligence,
King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia
Email: alfy@kfupm.edu.sa (E.-S.M.E.-A.); (A.S.S.); (O.A.A.-O.); (A.M.A.); (S.A.-A.)

*Corresponding author

Abstract—Recognizing human emotional and physical states using physiological measurements has several potential interdisciplinary applications and use cases across multiple domains, including detecting deceptive behaviors, employee screening, safety and security, mental health assessment, stress monitoring, and identifying learning difficulties and attention disorder. This paper presents and evaluates a novel solution approach for lie detection using multi-frame video models and convolutional neural networks for video processing. Different architectures have been evaluated and compared for detecting facial and bodily clues to enhance the detection accuracy. The video stream is preprocessed by selecting key frames and applying human face and body extraction techniques as inputs to the feature learning and detection model. In an ablation study, we also evaluated various ways to combine the trained face and body models under four scenarios: combining outputs of the same model architecture for both face and body, combining different model architectures for faces, combining different model architectures for body, and combining all model architectures for face and body. Evaluations using the Real-Life Trials dataset demonstrate the effectiveness of MobileNetV2 trained on full-body images, outperforming other test methods with more than 94% accuracy while significantly reducing computational costs and model size. Moreover, combining facial and bodily cues enhances model accuracy compared to using each modality in isolation.

Keywords—deception detection, lie detection, machine learning, deep learning, convolutional neural networks, physiological measurements

I. INTRODUCTION

There has been an increasing interest in automatic lie or deception detection methods as it is very crucial in many fields; in particular for law enforcement and crime investigation, national and personal safety, medical providers, insurance companies, and other businesses [1].

In the law enforcement field, deception detection can help police officers and authorities during crime investigation to detect liars while interrogating criminals. For national security measures, deception detection can help in protecting countries by screening potential employees, improving border control by early detection of foreign adversaries and suspicious travelers, identifying potential espionage and sabotage threats, and deterring unauthorized disclosure of classified information. In the medical field, it can help doctors and therapists obtain truthful information during interventions with patients as the latter may hide information that is critical to their lives; thus, providing more accurate diagnosis and treatment plans and maintaining proper health records for legal and insurance purposes. Also, businesses can benefit from the deception detection systems during job interviews, e.g. to ensure that applicants properly fit in the target position, especially if the position is critical or sensitive. Although there is no universally agreed-upon definition of the term deception, it is widely recognized through its key elements as described by psychologist Aldert Vrij, who has written extensively on the subject. He defined deception as “a successful or unsuccessful deliberate attempt, without forewarning, to create in another a belief which the communicator considers to be untrue” [2]; i.e., it involves a purposeful or deliberate intent to make someone believe something false without prior warning.

The polygraph test, one of the oldest well-known methods, was invented in 1921 by John Augustus Larson at Berkeley Police Department. It works by measuring and analyzing psycho-physiological changes in the human body associated with arousal and emotional activities to determine whether a person is lying [1]. The examination process involves questioning the subject while simultaneously measuring changes in autonomic and somatic responses by indicators such as changes in heart rate, blood pressure, respiratory patterns, and skin conductance due sweat gland activity to control the body

Manuscript received September 24, 2024; revised October 30, 2024; accepted January 28, 2025; published April 16, 2025.

temperature. The underlying principle is based on a subjective conjecture that deceptive responses produce physiological reactions that differ from those associated with non-deceptive responses. However, the polygraph test has several limitations including lack of theoretical foundation or empirical evidence, significant variation among individuals, and nonexistence of unique psychological responses exclusively associated with deception¹. Thus, traditional polygraph tests are often unreliable albeit being easy to use. Although the automation of these tests can increase their effectiveness, they still cannot detect lies with very high accuracy, and their results can be affected by many other factors including medical conditions such as asthma, Parkinson's disease, hypertension, and depression [3, 4].

With the revolution of artificial intelligence and machine learning, the problem of lie detection has attracted a growing interest within the research community [5]. The concept has even evolved with new approaches introduced to detect lies using various modalities such as audio, text, facial expressions, hand gestures, and body language. It has also been extended to include other deceptive digital content and fake news on social media platforms [6, 7]. Additionally, physiological methods have been developed to measure body movements, including facial muscle movements and micro-expressions as well as using sensors and thermal imaging to detect changes in body temperature during truth-telling or lying. Deception cues can be classified into verbal and non-verbal indicators, each encompassing various subcategories [8, 9]. Verbal cues are extracted from written or spoken text analysis whereas non-verbal cues are extracted from several other physiological measures such as full-body motion, head movement, facial expressions, eye gaze, pupil dilation, and eye blinking [10].

Recent research has increasingly emphasized non-verbal cues, particularly facial expressions, which are typically categorized into macro-expressions and micro-expressions based on their duration [9]. Macro-expressions occur more frequently and last longer (e.g., 0.5 to 4 s), therefore they are easier to control. On the other hand, micro-expressions are brief, subtle, localized facial movements that are less likely to be perceived by human eye. They often arise from failed attempts to hide or suppress emotions [11]. Though micro-expressions, which are involuntary generated by facial muscle movements, can serve as crucial cues for detecting deception, their reliability and practical utility for high stakes security screening remain a subject of debate. Recent studies have highlighted several limitations including infrequent occurrence in real-world scenarios and their presence can inconsistently happen whether a person is lying or not depending on situation [12]. Given these limitations, our research study aims at comparing the performance of different convolutional neural networks for detecting lies based on a broader range of facial and bodily characteristics extracted from multi-

frame visual models of video data. Additionally, our study includes an ablation analysis to assess the effectiveness of fusing various models trained on face and body modalities to address the limitation of each modality and potentially improve the accuracy and reliability of the automated lie detection approach.

The remainder of the paper is organized as follows. Section II presents a brief review of related work and discussion of ethical concerns. Section III explains the proposed methodology. Section IV describes the experimental work and discusses the results. Section V concludes the paper.

II. RELATED WORK

Deception detection is a complex task that has attracted a lot of attention over years. It requires monitoring certain behavioral and psychological aspects. Several approaches have been proposed in the literature based on polygraph tests and machine learning [5, 13, 14]. In this section, we briefly review work related to machine and deep learning using facial expressions, eye tracking, voice stress analysis, and brain imaging. In recent years, machine and deep learning algorithms have been intensively explored with great success for deception detection based on natural language processing, signal processing, and computer vision [13–16].

A number of studies have been proposed using facial analysis to detect deception [9]. In these studies, facial characteristics are captured from images and videos, including macro/micro expressions, eye gaze, eyebrows, eye blinks, lip movement, and head postures. For instance, Monaro *et al.* [17] proposed a model based on facial micro-expression features to detect deception. Their solution used a dataset consisting of 62 videos of Italian participants (43 females and 19 males), where 32 were truth-tellers and 30 acted as liars. The participants were interviewed in three phases: baseline phase where the interviewee presents his/her autobiographical data, free speech phase where the interviewee freely talks about a past real/fake holiday, and unexpected questions and answers phase. The feature extraction was conducted using three methods: Handcrafted features using Improved Dense Trajectory (IDT), 3D Convolutional Neural Network (C3D), and OpenFace to extract high level features and Action Units (AUs). Three different detection methods have been used: Support Vector Machine (SVM), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN). The authors found that SVM with OpenFace has the best performance with highest area under the curve (AUC) ranging from 0.72 to 0.78.

Tracking changes in eye movement patterns such as changes in pupil dilation, looking direction, and looking duration have been used as signs of deception in [18]. Eye-tracking technology can be used in real time, making it useful in situations that require immediate feedback. As an example, a real-time deception detection approach that integrates fine grained features of eyes and facial micro-movements is described in [19]. However, it requires special equipment and may not be suitable for every

¹ <https://www.apa.org/topics/cognitive-neuroscience/polygraph>

situation. Additionally, eye movements may be affected by the environment in which they occur, leading to erroneous results.

A number of research works have also considered additional physiological modalities such as pulse rate, heart rate, respiration rate, skin temperature, and brain waves. For example, Tsuchiya *et al.* [20] combined facial expressions with pulse rate. Brain imaging techniques such as Functional Magnetic Resonance Imaging (fMRI) and Electroencephalography (EEG) have been used to investigate lying behavior [21]. These techniques measure changes in brain activity when a person lies. For people who are good at lying or have training in lying, it will not be very useful in detecting lies. It also requires special equipment and may not be suitable for every situation. Another approach based on EEG signals and DWT decomposition has been presented in [22] and achieved an accuracy of more than 97%.

Automatic lie detection resources are generally difficult to collect and find due to the real-life nature of the samples to reflect deceptive behaviors in real scenarios. Also, the amount of data in available resources or datasets are limited as it is labor intensive task which requires a well-structured procedure and environment, while taking into consideration the challenge of ethical aspects and how many participants we can secure in each experiment. Last but not the least, the modality of the dataset (video, audio, text, EEG, physiological, eye gaze, etc.) is critical as it limits the subsequent modeling approach if the research is targeting to build a multimodal model. One of the most used datasets in lie detection is the real-life trials which consists of videos collected from court trials [23]. The dataset is multimodal and supports three modalities: video, audio, and text. The dataset consists of 121 videos (61 deceptive and 60 truthful trials). Videos in the dataset are 28.0 s on average. The participants were total 56 (21 females and 35 males), with the ages range between 16 and 60 years. The dataset was labeled as truthful or deceptive considering three scenarios or outcomes of the trial: guilty verdict, non-guilty verdict, and exoneration. for non-verbal attributes (mainly facial expressions and hand gestures), the annotation was carried out based on MUMIN scheme using Elan Software. At the video level, annotators put one label per each of the gestures (general face, lips, eyebrows, eyes, gaze, mouth, head movements, hand, and hand trajectory). This dataset has been utilized in a number of studies including [11, 23–27].

Some previous work focused on integrating verbal and non-verbal modalities including text, audio and vision for lie detection [11]. Using text and audio with RNN or LSTM has achieved an accuracy ranging from 76% to 84% while using micro-expression approach has led to a better accuracy in the range 77-88%. Another approach was proposed by Perez-Rosas *et al.* [23] using Random Forest and Decision Tree algorithms to analyze text and videos and achieved an accuracy of 75%. When including SVM, the accuracy reached 82% [24]. The dataset used in their work was also used in [25] combining text, voice and video to boost the accuracy to 96%. A higher

accuracy of 97% was obtained by Chebbi and Jebara [26] using the real-life trial data to train a KNN classifier with the three modalities of text, voice and video.

Gupta *et al.* [28] introduced another dataset, known as Bag-of-Lies. It incorporates multiple modalities including EEG and Gaze data with traditional video and audio data. The dataset was collected from 35 participants specifically for the automated deception detection based on realistic scenario. In total, the dataset contains 325 manually annotated recordings, comprising 162 lies and 163 truths. Each video recording varies in length from 3.5 s to 42 s. During data collection, each participant was shown between 6 and 10 images, one at a time, and then asked to describe each image either honestly or deceptively.

Some other datasets, such as the one utilized in [17], were collected in hypothetical environments by interviewing the participant, asking questions about specific subject and capture their responses accordingly in each experiment category (truth-telling, or lying). Another similar dataset was explored known as the avatar boarder agent interview dataset [29]. The interview revolves around airport security scenarios, where 100 participants interviewed (through 12 questions) by machine-based avatar agent about packing a suitcase and going to the airport for a holiday. Participants were divided equally into two truthful and deceptive scenarios. A total of 1200 short video clips, with an average video duration between 3 and 6 min, was recorded (one video per question, total 12 questions per participant).

While the results of machine learning based lie detection are encouraging, there are several critical ethical concerns that require careful consideration while developing a sophisticated lie detection system [16]. Among these issues are: (1) the potential misuse of data and systems, which could infringe on personal freedom and expose individuals to identity theft or other harmful uses of personal data, (2) the protection of human privacy and autonomy, ensuring individuals have the right to control over their personal data and the contexts in which they are used, (3) the risk of model bias and false positives as machine learning models can inadvertently mislabel trustful individuals as deceptive, particularly if they struggle with nuanced human behavior or trained on biased datasets; leading to unjust outcomes and higher stress for those falsely accused, and (4) limited explainability, as black-box models make it difficult to clarify how decisions are reached or to provide recourse when errors occur.

III. METHODOLOGY

In this study, a series of experiments were designed and developed to evaluate and compare different convolutional neural networks for lie detection, using facial and body spatiotemporal features learned from the video stream. Fig. 1 illustrates the overall workflow considered in this study. After the model is trained, it is evaluated on a separate unseen dataset. The trained model is also used for real time inference through a custom-developed web-based application.

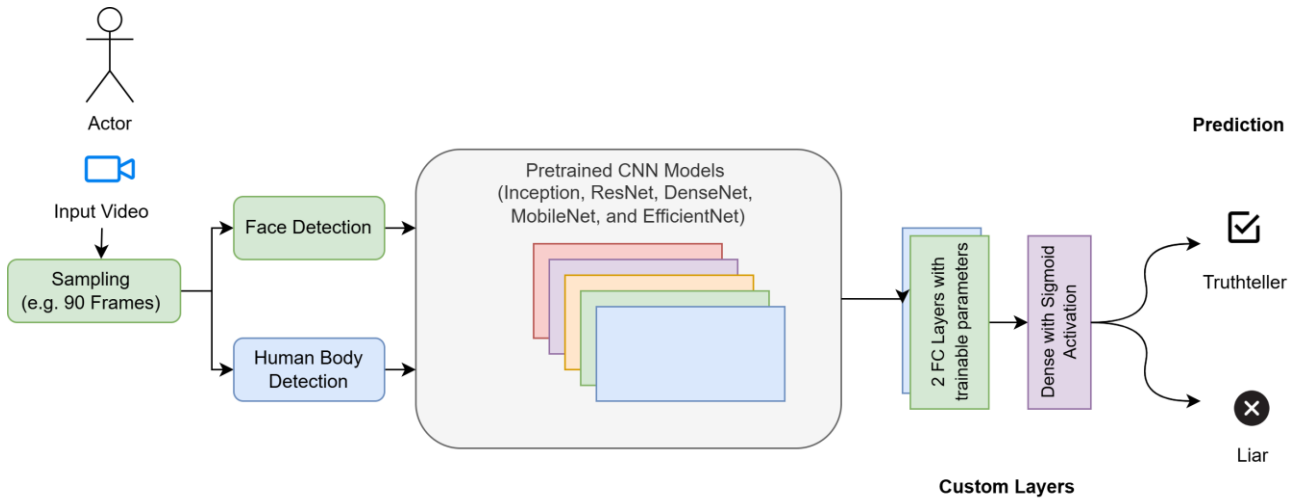


Fig. 1. Workflow of the proposed lie detection system.

A. Video Data Collection and Preparation

- Video files are first collected and prepared for subsequent processing. the entire dataset is split into train, validation and test sets. Each video is composed of several frames and the number of frames depends on the resolution and length of the video. In the adopted dataset, the average number of frames per second is approximately 30 while the average video length is approximately 30 s or a total of 90 frames per video on average. Rather than processing all such frames, which will require high computational resources, we limited the number of frames in two ways as follows:
- Selecting frames by uniform sampling throughout the video and setting the interval based on the total video frames and the maximum frames the model will take as input. To avoid noisy frames, in our experiments we have ignored the first 10% and last 5% of frames in the video. Although this is a straightforward approach, it may miss key frames with information content related with deception and subsequently decrease the results.
- Selecting the frames based on frame differencing to capture motion by comparing adjust frames for pixel-wise changes. The idea is to compute the absolute difference between each frame and its previous one in the sequence, then compare it to a threshold value to obtain a motion mask.

To improve the practical deployment of the proposed system for scenarios where the video camera may not be fixed or recording quality is not high, several preprocessing methods can be applied, e.g. camera motion compensation and image stabilization algorithms, adaptive masking that can adjust to changes in camera positions and angles or subject orientation, image quality enhancement to address shadow effect, blurring, and noise, and lighting normalization to mitigate the effect of varying illumination. Other preprocessing techniques may include augmentation to increase the size the dataset or balance the class distribution, which can generally

improve the model performance. However, unlike machine learning for other tasks, synthetic artifacts have limited utility in lie detection since they can reduce the subtle cues related to genuine human signals related to lie detection leading to non-authentic psychological or physiological responses and risk of overfitting to the synthetic artifacts. Therefore, we avoided using synthetic samples in order not to change the nature of the original dataset and have fair comparison among different models.

B. Face/Body Segmentation

Since Haar cascade face detector [30] is traditional, it is used as a baseline, and MTCNN, which is a cascading series of CNNs, is used to detect, align, and extract facial features from digital images with high accuracy [31–33]. In the preprocessing stage, we have used and compared results of both approaches on a sample from the dataset. Even though MTCNN showed a slightly better accuracy on some samples, its performance was similar for the majority of samples. Also, MTCNN took longer to extract regions of interest, making it slower for real-time deployment.

Besides, we have implemented human detection using Detectron2, which is an open-source object detection and segmentation framework built on PyTorch. It was built by Facebook AI Research (FAIR) to support rapid implementation in the computer vision field and has many algorithms; we have selected Mask R-CNN for body segmentation. This approach will likely capture more movements of the entire body, head poses, hand gestures, etc. while the subject is speaking, helping to improve the features extracted from videos to make more accurate predictions.

C. CNN Network Architectures

After experimenting with several model architectures and pre-processing methods, the final model design is chosen to be based on MobileNetV2 as a pre-trained model with tweaking of some of hyperparameters of initial design. The main reason behind using MobileNetV2 is due to the efficient and effective use of this model in video classification task, making it suitable

for deployment without sacrificing on the performance of the model. Also, we added human detection beside face detection during the pre-processing stage. Below is a brief description of the evaluated architectures:

1) *Inception*: This architecture was introduced in [34] to allow multi-level feature extraction by computing convolutions with different kernel sizes (1×1 , 3×3 , 5×5), which were then stacked in the channel dimension and fed to the following layer in the network. Inception V3 also has fewer parameters than VGGNet and ResNet.

2) *ResNet*: Another architecture considered in this study is ResNet [35], which was developed to alleviate the limitations of the traditional CNNs through providing skip connections to bypass the non-linear transformation and create an identity mapping. This contributes to improving the results, enhancing the scalability, and reducing the complexity through learning residuals between inputs and outputs. There are different versions of ResNet based on the number of layers, e.g. ResNet-50 with 50 layers. The architecture begins with an initial convolutional layer using a 7×7 kernel with 64 filters and a stride of 2, followed by a 3×3 max-pooling layer with a stride of 2. It then features four stages of residual blocks.

3) *DenseNet*: In typical CNNs with L sequential layers, there is one main forward connection from layer i to layer $i+1$ for $i = 1$ to L , yielding a total of $L-1$ direct connections. In contrast, ResNet employs residual skip connections as well, resulting in $2(L-1)$ connections. DenseNet [36], however, introduces a more complex architecture where each layer is directly connected to all subsequent layers, leading to a total of $L(L-1)/2$ connections. Distinct from ResNets, which use residual mappings to enhance information flow, DenseNets concatenate the outputs of previous layers to preserve and integrate information more effectively throughout the network. In this study, we considered DenseNet121 in our experiments.

4) *MobileNet*: MobileNets [37] are efficient models optimized for mobile and embedded vision applications. They use depth-wise separable convolutions to build lightweight neural networks. With two global hyper-parameters, MobileNets allow for an effective balance between latency and accuracy, making them adaptable to different application constraints. The main components of a MobileNet are:

- **Depthwise Separable Convolution**: which consists of a depth-wise convolution followed by a point-wise 1×1 convolution. The former applies a single filter to each input channel while the latter combines the outputs from the depth-wise convolution.
- **Initial Layers**: The network starts with a 3×3 convolutional layer with a stride of 2, followed by batch normalization and ReLU activation.
- **Residual Blocks**: MobileNet employs a series of depthwise separable convolutional blocks arranged in a sequence to build the network.
- **Global Average Pooling and Fully Connected Layer**: At the end of the network, global average pooling is used to reduce the feature map to a single value per channel, followed by a fully connected layer to produce the final classification output.

Different variations have been considered in our study including MobileNetV2, MobileNetV3Small, and MobileNetV3Large.

5) *EfficientNet*: EfficientNetV2 [38] introduced another convolutional network family that offers faster training and better efficiency. By combining neural architecture search and scaling with novel operations like Fused-MBConv, these models are up to 6.8 times smaller and train significantly faster. To mitigate accuracy drops from increasing image sizes, adaptive regularization techniques are used. There are different versions of EfficientNetV2. In this study, we experimented with the small version, EfficientNetV2S, in addition to EfficientNetV2B0.

D. Model Deployment, Fusion, and Tuning

Fig. 2 shows various workflows for training face and body models. The output scores S_F and S_B are thresholded to decide based on the face only $D_F = h_\theta(S_F) \in \{0, 1\}$ or body only $D_B = h_\theta(S_B) \in \{0, 1\}$, where h_θ is a threshold function. Additionally, we combined the output at the score level and at the decision level. Scores are combined by taking the average or weighted average based on the performance of each model on separate modalities. For example, for face $S = (S_F + S_B)/2$ or $S = (W_F \cdot S_F + W_B \cdot S_B)/(W_F + W_B)$, for unweighted and weighted average, respectively, where W_F and W_B are the corresponding weights based on each model performance.

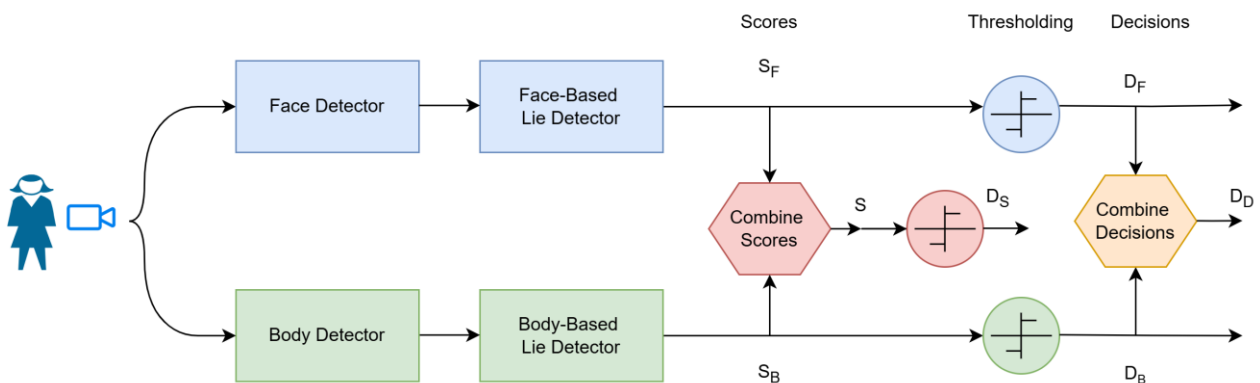


Fig. 2. Various paths to deploy and combine separate models trained on face and body modalities (blue path for face and green path for body).

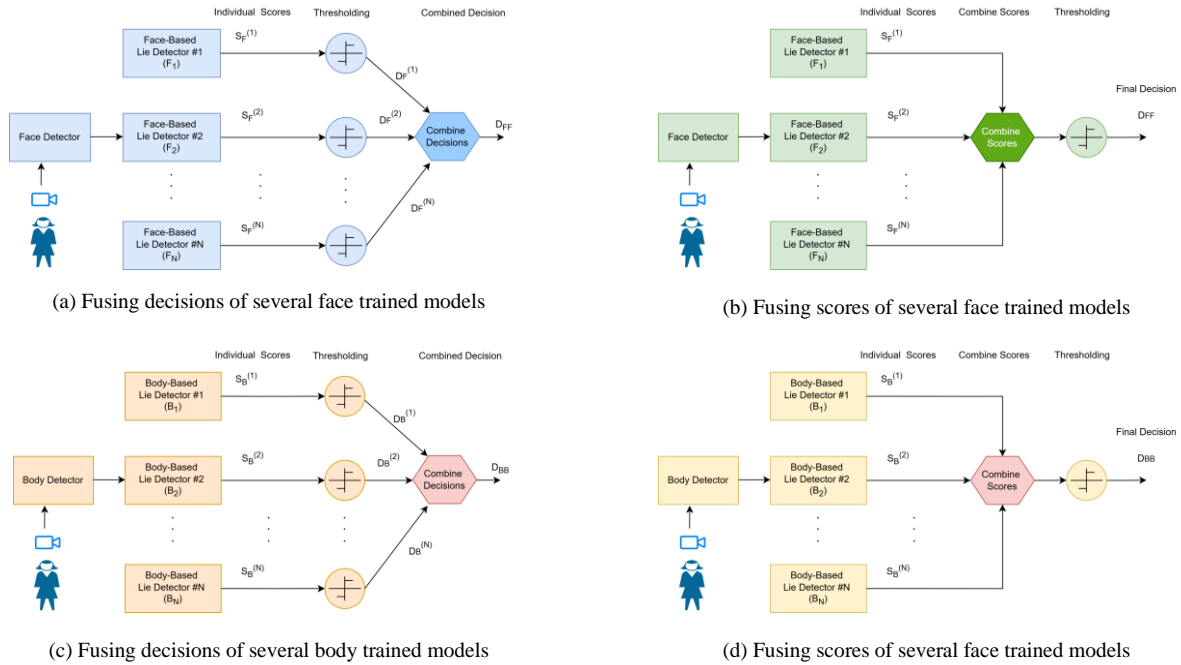


Fig. 3. Fusing multiple pretrained models on faces and body modalities.

To combine decision, the final decision is made based on the coincidence of the models' outputs and ties are resolved randomly since there are only two modalities. Furthermore, in an ablation study, we tested other ways to combine the outputs of the trained models on faces and bodies as shown in Fig. 3. Several models trained on facial expressions are combined in two different approaches: (a) majority vote after thresholding individual model scores (as illustrated in Fig. 3 (a)), and (b) averaging the scores of all models then thresholding to get the final decision (as illustrated in Fig. 3 (b)). Similar fusion techniques are also applied to the body modality models as shown in Fig. 3 (c) and (d), respectively. The configuration of CNN hyperparameters can significantly impact model performance. Although systematic tuning methods—such as grid search, random search, Bayesian optimization, or Hyperband—can be employed, these approaches introduce additional computational overhead during training and may prevent fair comparisons across different models.

E. Real-Time Lie Detector

A feature has been introduced into our solution which is real-time deception detection by capturing a real-time stream from the user's camera. Two options were tested for real-time face-detection: OpenCV's Haar Cascade face detection and Multi-Task Cascaded Convolutional Networks (MTCNN). While MTCNN is considered a more effective modern method compared to Haar Cascade, it failed in our implementation from the speed perspective. Haar showed to be faster in combining the frames and predicting the probability despite a smaller lag in the video stream (but still faster than MTCNN). Hence, Haar Cascade has been chosen to detect faces in our real-time lie-detection system. Once the face is detected in the real-time environment, a square box will

be drawn around it, and the system will keep capturing 20 frames and doing the prediction based on them until the user ends the stream. A text string is displayed in green/red color showing "Trustful" or "Liar" together with a confidence score. The user will experience some lag in the video due to the continuous operation of capturing 20 frames and doing the prediction, but still reasonable.

F. Web Application

The complete lie detection solution has been developed as a web application using Flask, which is a lightweight web framework for developing Python applications on the web. Fig. 4 presents sample snapshots of the user interface and results. The interface allows the users to either upload a video or utilize the system web or mobile camera for real time processing. To enhance the user experience and visual appeal, the solution's design was implemented using Bootstrap and CSS frameworks, resulting in a modern and intuitive easy-to-use interface. The best model will be incorporated into the solution to give the best accuracy for lie detection process. This application has two use cases:

- Upload a video file to be analyzed: It allows a user to upload a video file in MP4 format to be analyzed by the system. The system analyzes the video and detects whether the speaker is telling lies or truth based on confidence score returned by the model. If the score is 50 or above, this means "lie", otherwise, it is "truth".
- Real-time detection: It allows a user to open a mobile or web camera and try the system on real-time basis. The system will detect his face and display the score of truth or lie while the speaker is talking.

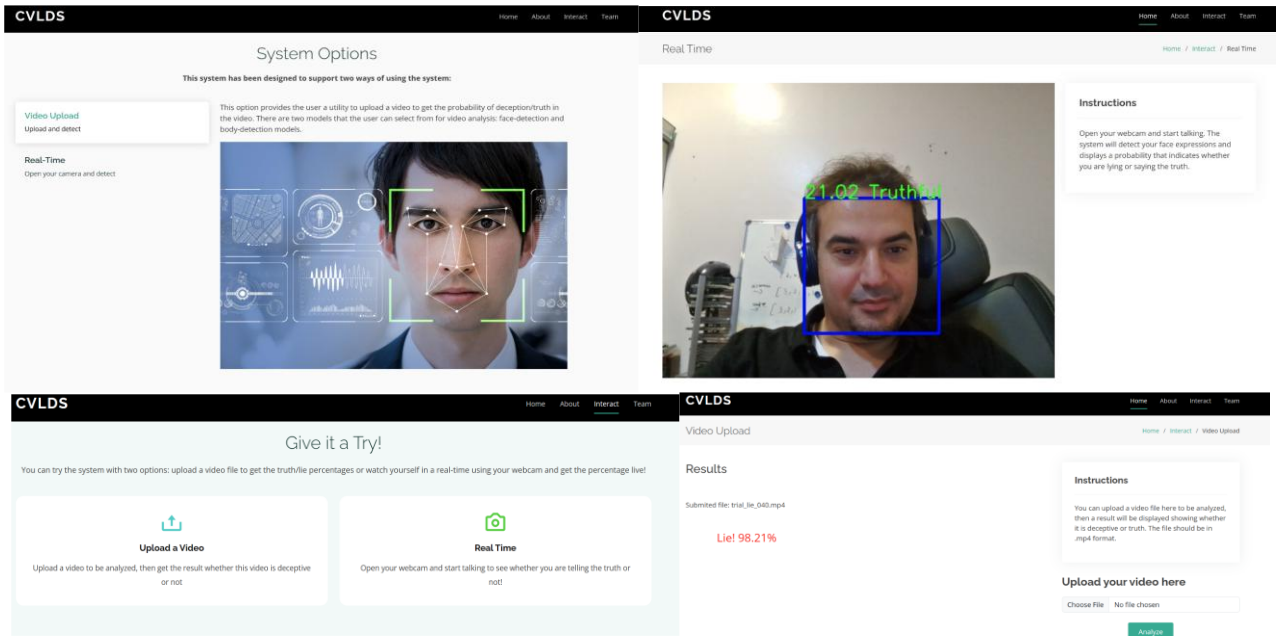


Fig. 4. Fusing multiple pretrained models on faces and body modalities. Web-based application and for uploading videos and detecting deception; also allows real time inference on data captured using web or mobile camera.

IV. EXPERIMENTAL WORK AND EVALUATION

A. Experimental Settings

For developing and training the deep learning models, we have developed our code in Python as Jupyter notebook and conducted the experiments on Google Colab using A100 GPU and 32 GB RAM. We conducted several experiments to evaluate the proposed approach on a benchmark dataset using different performance metrics in terms of model accuracy, precision, recall and F1 score. We also considered the model size and number of parameters. Furthermore, we have tested the detection model (for face) since it is a lightweight model on regular laptop of specs (i7 Quad Core CPU 3.30 GHz, 32 GB RAM, GPU: RTX3050 Ti) and the performance was acceptable. However, for body-based model, the same specifications were not enough to have smooth streaming and detection. We did one experiment for CPU versus GPU during inference for body model, prediction result with GPU was faster almost five times faster.

B. Dataset

In our experiments, we adopted the Real-Life Trials dataset as one of the most popular datasets in lie and deception detection [23, 39]. It consists of 121 videos collected from court trials available through public multimedia sources; out of them 61 videos are deceptive and 60 videos are truthful trials. Each video in the dataset is 28s on average. It is collected from a total of 56 participants (21 females and 35 males), with ages ranging between 16 and 60 years. The dataset was labeled as truthful or deceptive considering three scenarios or outcomes of the trial: guilty verdict, non-guilty verdict, and exoneration. The dataset has multiple modalities

allowing research on verbal and non-verbal cues using visual, auditory and textual attributes. The annotations of the dataset were carried out based on MUMIN multimodal coding scheme using Elan Software [40]. At the video level, annotators assign one label per each of the gestures (general face, lips, eyebrows, eyes, gaze, mouth, head movements, hand, and hand trajectory). This dataset is available for education and research purposes only and can be downloaded from the University of Michigan². In this work, we focused on the visual modality through facial expressions and body postures. Fig. 5 shows sample images of the dataset.

C. Results and Discussions

We conducted extensive experiments to assess different CNN architectures on the Real-Life Trials dataset. We begin by preprocessing the dataset, removing noisy video files that did not clearly display the target subject's face during speech in the court proceedings. This filtering process resulted in a refined dataset of 111 videos, which are then partitioned into 77 videos (69.4%) for training, 17 videos (15.3%) for validation, and 17 videos (15.3%) for testing.



Fig. 5. Sample images from the Real-Life Trials dataset.

² <https://public.websites.umich.edu/~zmohamed/resources.html>

Additionally, for face detection, we experimented with two preprocessing approaches for face detection: OpenCV’s Haar Cascade, and Multi-task Cascaded Convolutional Networks (MTCNN). For frames selection, we ran a pilot test for two approaches: frame differencing and frames sampling; and ended up using frames sampling. Experiments were conducted using the following eight CNN architectures: DenseNet121, EfficientNetV2B0, EfficientNetV2S, InceptionV3, MobileNetV2, MobileNetV3Large, MobileNetV3Small, and ResNet50. They are evaluated under different scenarios on body and face modalities. The performance is assessed on the test dataset in terms of model accuracy, precision, recall, F1 score. We also take into consideration the model size and number of parameters.

In Table I, the results are reported for the developed lie detection models on face and body modalities when they are used separately in terms of precision, recall, F1 score and accuracy on the test dataset. We found that MobileNetV3Small achieved the highest accuracy of 88.24% among all models trained on faces. On the other hand, when using the body modality, MobileNetV2 has the highest overall accuracy of 94.12% and EfficientNetV2B0 has the second highest accuracy of 88.24%. Moreover, InceptionV3 has the lowest accuracy

in both face and body modalities. When comparing all models for face and body modalities, it is observed that the body modality has only better accuracy in two cases when using EfficientNetV2B0 or MobileNetV2; this can be attributed to the fact that the face expressions and muscle changes can reveal more insights for lie detection. Fig. 6 shows the performance in terms of test accuracy, total number of parameters, and size in memory for each model on both body and face modalities. As MobileNetV2 has the highest accuracy with reasonable size, it will be a good choice, especially for resource-constrained devices.

Additionally, the scores for face and body are combined: by treating them equally and by weighting the scores by the model performance for each modality alone, the details are presented in the methodology section. The results are shown in Table II. It is observed that the performance for most cases is better for weighted average than separate modalities or simple average of scores. We also tested two other scenarios: decision level and score level fusions. For decision levels, based on the final decision is based on the majority of models. For score level, scores are combines for all face models and for all body models. The results are shown in Table III.

TABLE I. PERFORMANCE COMPARISON OF DIFFERENT MODELS ON THE TEST DATASET FOR FACE AND BODY MODALITIES

Model	Modality							
	Face				Body			
	Prc. (%)	Rec. (%)	F1 (%)	Acc. (%)	Prc. (%)	Rec. (%)	F1 (%)	Acc. (%)
DenseNet121	70.00	87.50	77.78	76.47	83.33	62.50	71.43	76.47
EfficientNetV2B0	77.78	87.50	82.35	82.35	87.50	87.50	87.50	88.24
EfficientNetV2S	71.43	62.50	66.67	70.59	71.43	62.50	66.67	70.59
InceptionV3	66.67	75.00	70.59	70.59	44.44	50.00	47.06	47.06
MobileNetV2	83.33	62.50	71.43	76.47	88.89	100.00	94.12	94.12
MobileNetV3Large	83.33	62.50	71.43	76.47	66.67	50.00	57.14	64.71
MobileNetV3Small	87.50	87.50	87.50	88.24	75.00	75.00	75.00	76.47
ResNet50	85.71	75.00	80.00	82.35	70.00	87.50	77.78	76.47

TABLE II. PERFORMANCE COMPARISON OF DIFFERENT MODELS ON THE TEST DATASET WHEN FUSING FACE AND BODY MODALITIES UNDER TWO DIFFERENT SCENARIOS: EQUALLY WEIGHTED OR WEIGHTED BY THE MODEL PERFORMANCE ON THE CORRESPONDING MODALITY (LAST TWO COLUMNS SHOW PERFORMANCE IMPROVEMENT OVER FACE AND BODY WHEN EACH IS USED ALONE)

Model	Modality									
	Face+Body				Weighted Face+Body					
	Prc. (%)	Rec. (%)	F1 (%)	Acc. (%)	Prc. (%)	Rec. (%)	F1 (%)	Acc. (%)	Imp Face	Imp Body
DenseNet121	71.43	62.50	66.67	70.59	83.33	62.50	71.43	76.47	0	0
EfficientNetV2B0	87.50	87.50	87.50	88.24	87.50	87.50	87.50	88.24	0.059	0
EfficientNetV2S	85.71	75.00	80.00	82.35	71.43	62.50	66.67	70.59	0	0
InceptionV3	50.00	50.00	50.00	52.94	66.67	75.00	70.59	70.59	0	0.235
MobileNetV2	83.33	62.50	71.43	76.47	88.89	100.00	94.12	94.12	0.176	0
MobileNetV3Large	80.00	50.00	61.54	70.59	83.33	62.50	71.43	76.47	0	0.118
MobileNetV3Small	87.50	87.50	87.50	88.24	87.50	87.50	87.50	88.24	0	0.118
ResNet50	87.50	87.50	87.50	88.24	85.71	75.00	80.00	82.35	0	0.059

TABLE III. PERFORMANCE COMPARISON ON THE TEST DATASET WHEN DECISION OR SCORE OF THE MODELS ON FACE AND BODY MODALITIES

Model	Modality							
	Decision				Score			
	Prc. (%)	Rec. (%)	F1 (%)	Acc. (%)	Prc. (%)	Rec. (%)	F1 (%)	Acc. (%)
Face	85.71	75.00	80.00	82.35	85.71	75.00	80.00	82.35
Body	77.78	87.50	82.35	82.35	85.71	75.00	80.00	82.35
Face+Body	85.71	75.00	80.00	82.35				
Weighted Face+Body	87.50	87.50	87.50	88.24				

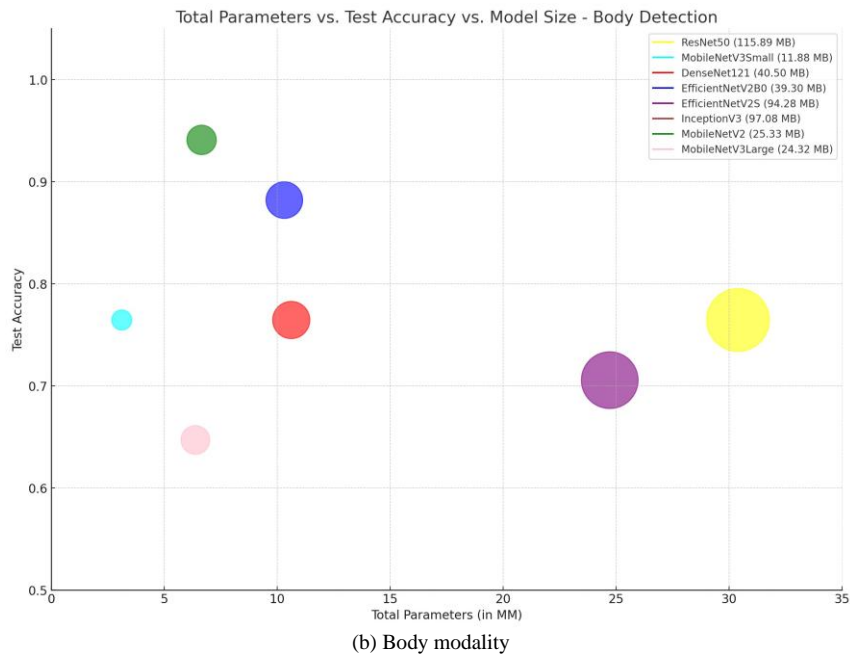
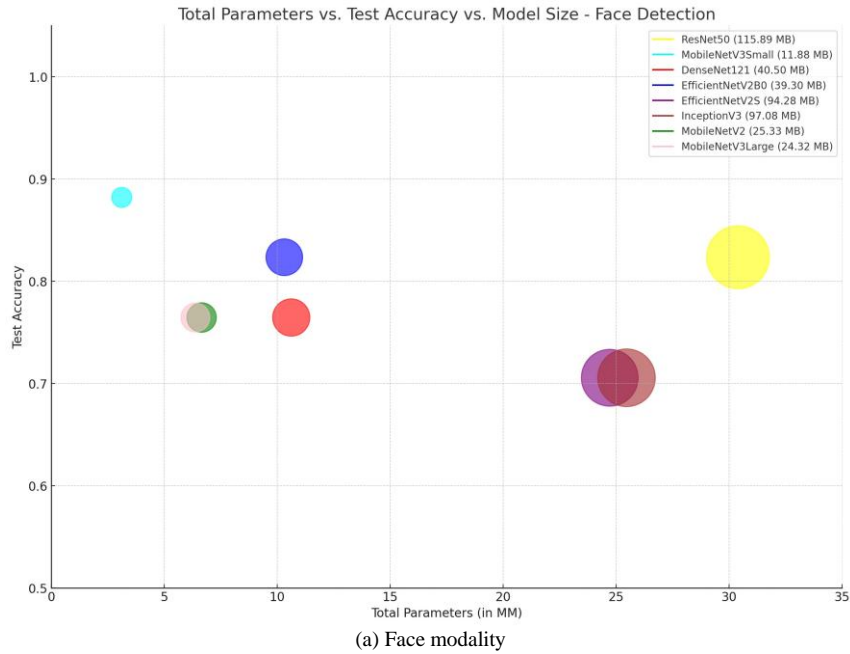


Fig. 6. Performance tradeoffs in terms of model accuracy, total number of parameters and memory size.

V. CONCLUSION AND FUTURE WORK

Automatic lie detection is a challenging problem, but has several potential applications for national security, law enforcement, crime investigation, banking and credit card companies as well as personal security and safety. This study investigates diverse CNN-based lie detection approaches by learning from visual data to predict deceptive behavior through facial micro-expressions and body posture changes. Various models have been developed and evaluated including DenseNet121, EfficientNetV2S, EfficientNetV2B0, InceptionV3, MobileNetV3Large, MobileNetV3Small, MobileNetV2 and ResNet50. The evaluation metrics include the

detection accuracies, model size, and number of parameters. The results show that using MobileNetV2 and MobileNetV3Small can achieve high accuracy while being ideal for deployment in resource-constrained devices due to relatively reduced sizes and number of parameters. Additionally, we have explored various methods to combine face and body modalities at different levels, addressing the inherent limitations of each when used in isolation. The experimental results demonstrate promising improvement of the model accuracy, emphasizing the need for a multi-faceted approach that could potentially enhance the robustness and reliability of automated deception detection. Some directions for future research include a more comprehensive approach

incorporating more modalities such as textual and audio modalities or any other data streams to bolster the model robustness and reliability, especially in the presence of different lighting conditions, occlusion and non-frontal faces, which will require other preprocessing methods to be applied. Another direction is to explore knowledge-informed machine learning to integrate domain specific knowledge and contextual information that further enhance model generalization and performance while being resilient to adversarial counterfeits.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

ASS with OAA and AMA drafted the paper, developed the experiments; ESM conceptualized and discussed the paper's contribution, conducted the model fusion, finalized the paper writing, visual illustrations, managed the team, submitted the paper and managed the revision; SAA with ASS, OAA and AMA conducted the first experiments and prepared a draft of the work. All authors have participated in conceptualizing the idea and proofread the final version of the paper.

FUNDING

This research was funded by King Fahd University of Petroleum and Minerals, the Interdisciplinary Research Center for Intelligent Secure Systems (IRC-ISS) under Grant No. INSS2204. The team would also like to thank SDAIA-KFUPM Joint Research Center for Artificial Intelligence for collaboration during this work.

REFERENCES

- [1] D. Grubin and L. Madsen, "Lie detection and the polygraph: A historical review," *The Journal of Forensic Psychiatry & Psychology*, vol. 16, no. 2, pp. 357–369, 2005.
- [2] A. Vrij, "Deception and truth detection when analyzing nonverbal and verbal cues," *Applied Cognitive Psychology*, vol. 33, no. 2, pp. 160–167, 2019.
- [3] C. R. Honts and S. Amato, "Automation of a screening polygraph test increases accuracy," *Psychology, Crime & Law*, vol. 13, no. 2, pp. 187–199, 2007.
- [4] W. M. Waid and M. T. Orne, "The physiological detection of deception: The accuracy of polygraph testing can be affected by such variables as attentiveness, drugs, personality, and the interaction between examiner and subject," *American Scientist*, vol. 70, no. 4, pp. 402–409, 1982.
- [5] A. Constancio, D. Tsunoda, H. Silva, J. Silveira, and D. Carvalho, "Deception detection with machine learning: A systematic review and statistical analysis," *PLOS One*, vol. 18, no. 2, e0281323, 2023.
- [6] H. Alaskar, Z. Sbat, W. Khan, A. Hussain, and A. Alrawais, "Intelligent techniques for deception detection: A survey and critical study," *Soft Computing*, vol. 27, no. 7, pp. 3581–3600, 2023.
- [7] K. Qureshi, R. Malick, M. Sabih, and H. Cherifi, "Deception detection on social media: A source-based perspective," *Knowledge-Based Systems*, vol. 256, 109649, 2022.
- [8] D. Kucuk and F. Can, "Stance detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 1, pp. 1–37, 2020.
- [9] H. Delmas, V. Denault, J. K. Burgoon, and N. E. Dunbar, "A review of automatic lie detection from facial features," *Journal of Nonverbal Behavior*, vol. 48, no. 1, pp. 93–136, 2024.
- [10] S. Venkatesh, R. Ramachandra, and P. Bours, "Robust algorithm for multimodal deception detection," in *Proc. IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019, pp. 534–537.
- [11] Z. Wu, B. Singh, L. Davis, and V. Subrahmanian, "Deception detection in videos," in *Proc. the AAAI Conference on Artificial Intelligence*, 2018, vol. 32, no. 1.
- [12] S. Jordan, B. Laure Brimbal, D.B. Wallace, S. M. Kassin, M. Hartwig, and C. N. Street, "A test of the micro-expressions training tool: Does it improve lie detection?" *Journal of Investigative Psychology and Offender Profiling*, vol. 16, no. 3, pp. 222–235, 2019.
- [13] S. Azhan, A. Zaman, and M. R. Bhuiyan, "Using machine learning for lie detection: Classification of human visual morphology," Ph.D. dissertation, BRAC University, 2018.
- [14] S. V. Fernandes and M. S. Ullah, "A comprehensive review on features extraction and features matching techniques for deception detection," *IEEE Access*, vol. 10, pp. 28233–28246, 2022.
- [15] R. J. Dhabarde, D. Kodawade, and S. Zalte, "Hybrid machine learning model for lie-detection," in *Proc. IEEE 8th International Conference for Convergence in Technology (I2CT)*, 2023, pp. 1–5.
- [16] S. A. Prome, N. A. Ragavan, M. R. Islam, D. Asirvatham, and A. J. Jegathesan, "Deception detection using ML and DL techniques: A systematic review," *Natural Language Processing Journal*, 100057, 2024.
- [17] M. Monaro, S. Maldera, C. Scarpazza, G. Sartori, and N. Navarin, "Detecting deception through facial expressions in a dataset of videotaped interviews: A comparison between human judges and machine learning models," *Computers in Human Behavior*, vol. 127, 107063, 2022.
- [18] J. Zuo, T. Gedeon, and Z. Qin, "Your eyes say you're lying: an eye movement pattern analysis for face familiarity and deceptive cognition," in *Proc. IEEE International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–8.
- [19] W. Khan, K. Crockett, J. O'Shea, A. Hussain, and B. M. Khan, "Deception in the eyes of deceiver: A computer vision and machine learning based automated deception detection," *Expert Systems with Applications*, vol. 169, p. 114341, 2021.
- [20] K. Tsuchiya, R. Hatano, and H. Nishiyama, "Detecting deception using machine learning with facial expressions and pulse rate," *Artificial Life and Robotics*, vol. 28, no. 3, pp. 509–519, 2023.
- [21] F. A. Kozel, K. A. Johnson, Q. Mu, E. L. Grenesko, S. J. Laken, and M. S. George, "Detecting deception using functional magnetic resonance imaging," *Biological Psychiatry*, vol. 58, no. 8, pp. 605–613, 2005.
- [22] M. Aslan, M. Baykara, and T. B. Alakus, "Lstmncp: lie detection from EEG signals with novel hybrid deep learning method," *Multimedia Tools and Applications*, vol. 83, no. 11, pp. 31655–31671, 2024.
- [23] V. Perez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proc. the ACM International Conference on Multimodal Interaction*, 2015, pp. 59–66.
- [24] V. Perez-Rosas, M. Abouelenien, R. Mihalcea, Y. Xiao, C. Linton, and M. Burzo, "Verbal and nonverbal clues for real-life deception detection," in *Proc. the International Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2336–2346.
- [25] M. Gogate, A. Adeel, and A. Hussain, "Deep learning driven multimodal fusion for automated deception detection," in *Proc. IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1–6.
- [26] S. Chebbi and S. B. Jebara, "Deception detection using multimodal fusion approaches," *Multimedia Tools and Applications*, vol. 82, no. 9, pp. 13073–13102, 2023.
- [27] B. Bicer and H. Dibeklioglu, "Automatic deceit detection through multimodal analysis of high-stake court-trials," *IEEE Transactions on Affective Computing*, vol. 15, no. 1, pp. 342–356, 2024.
- [28] V. Gupta, M. Agarwal, M. Arora, T. Chakraborty, R. Singh, and M. Vatsa, "Bag-of-lies: A multimodal dataset for deception detection," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 1–8.
- [29] J. O'Shea, K. Crockett, W. Khan, P. Kindynis, A. Antoniadis, and G. Boultaidakis, "Intelligent deception detection through machine-based interviewing," in *Proc. IEEE International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1–8.

- [30] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001.
- [31] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
- [32] H. Ku and W. Dong, "Face recognition based on mtcnn and convolutional neural network," *Frontiers in Signal Processing*, vol. 4, no. 1, pp. 37–42, 2020.
- [33] S. Al-Azani and E.-S. M. El-Alfy, "Enhanced video analytics for sentiment analysis based on fusing textual, auditory and visual information," *IEEE Access*, vol. 8, pp. 136843–136857, 2020.
- [34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.
- [35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [36] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.
- [37] A. G. Howard, "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint, arXiv:1704.04861, 2017.
- [38] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *Proc. 2021 International Conference on Machine Learning*, 2021, pp. 10096–10106.
- [39] M. U. Sen, V. Perez-Rosas, B. Yanikoglu, M. Abouelenien, M. Burzo, and R. Mihalcea, "Multimodal deception detection using real-life trial data," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 306–319, 2022.
- [40] L. Allwood, L. Cerrato, L. Dybkjaer, K. Jokinen, C. Navarretta, and P. Paggio, "The MUMIN multimodal coding scheme," *NorFA Yearbook*, pp. 129–157, 2005.

Copyright © 2025 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).