

SVC 2025: the First Multimodal Deception Detection Challenge

Xun Lin*
Great Bay University
Dongguan, Guangdong, China
lxlunxun@gmail.com

Xiaobao Guo*
Nanyang Technological University
Singapore
xiaobao001@e.ntu.edu.sg

Taorui Wang*
Great Bay University
Dongguan, Guangdong, China
25B351018@stu.hit.edu.cn

Yingjie Ma*
Great Bay University
Dongguan, Guangdong, China
muring2248@outlook.com


Jiajian Huang*
Great Bay University
Dongguan, Guangdong, China
jiajian_huang_cs@163.com

Jiayu Zhang*
Great Bay University
Dongguan, Guangdong, China
qmmcxm@protonmail.com

Junzhe Cao*
Great Bay University
Dongguan, Guangdong, China
caojunzhe@buaa.edu.cn


Zitong Yu†
Great Bay University
Dongguan, Guangdong, China
zitong.yu@ieee.org

Statement read by the speaker - *"I have passed a toilet roll under the cubicle wall to Madonna!"*



Opponent

- Question 1** - Few sheets or the whole roll?
- Question 2** - How are you sure that person is Madonna? You cannot see the person from your cubicle!



Speaker

- Response 1** - I passed 5 sheets!
- Response 2** - I went to the nightclub 'Chinawhite' and heard that Madonna would be there. A lady next to my toilet cubicle asked me for a few toilet papers and later when I was washing hands, I saw Madonna pass by!

Opposing team's prediction: *"I think it's a Lie!"*
Veracity of the statement: *False*
Result: *Opposing team wins*



Figure 1: Examples of deceptive actions. The left side shows a game show scenario from Guo et al., illustrating a typical deceptive round. The right image displays selected frames showcasing truthful and deceptive behaviors from video clips provided by Soldner et al..

Abstract

Deception detection is a critical task in real-world applications such as security screening, fraud prevention, and credibility assessment. While deep learning methods have shown promise in surpassing human-level performance, their effectiveness often depends on the availability of high-quality and diverse deception samples. Existing research predominantly focuses

on single-domain scenarios, overlooking the significant performance degradation caused by domain shifts. To address this gap, we present the SVC 2025 Multimodal Deception Detection Challenge, a new benchmark designed to evaluate cross-domain generalization in audio-visual deception detection. Participants are required to develop models that not only perform well within individual domains but also generalize across multiple heterogeneous datasets. By leveraging multimodal data, including audio, video, and text, this challenge encourages the design of models capable of capturing subtle and implicit deceptive cues. Through this benchmark, we aim to foster the development of more adaptable, explainable, and practically deployable deception detection systems, advancing the broader field of multimodal learning. By the conclusion of the workshop competition, a total of 21 teams had submitted their final results. Our baseline is released at MMDD2025.

*Equal Contribution.

†Corresponding Author

CCS Concepts

• **Computing methodologies** → **Computer vision**.

Keywords

multimodal deception detection, cross-domain, generalization

ACM Reference Format:

Xun Lin, Xiaobao Guo, Taorui Wang, Yingjie Ma, Jiajian Huang, Jiayu Zhang, Junzhe Cao, and Zitong Yu. 2025. SVC 2025: the First Multimodal Deception Detection Challenge. In *Proceedings of (ACM Multimedia '25)*. ACM, New York, NY, USA, 6 pages. <https://doi.org/XXXXXX.XXXXXXX>

1 Introduction

Deception detection plays a crucial role in accurately assessing truthfulness and identifying deceptive behaviors, having pivotal applications in many fields such as credibility assessment in business, multimedia anti-fraud, and custom security [1, 5, 11]. With its significant intention, deception detection remains inherently difficult. As human bias towards assuming truthfulness, human precision remains around 54% [2], slightly above chance. To discover a better performance and a less time-consuming method, researchers have increasingly explored automated approaches that combine advances in computer vision, natural language processing, and deep learning for deception detection. Recently, deep learning methods have demonstrated their credibility, achieving comparable or surpassing human detection even in some complex tasks [3, 17, 18].

The performance of AI models in deception detection is heavily reliant on the availability of authentic and effective deception samples from the real world. While present models provide satisfactory results, fewer studies have explored the cross-domain issue, despite the presence of significant domain shifts in public deception detection datasets [9, 14, 16, 19]. The generalizability of the models is critical for practical applications. Therefore, such domain shifts need to be investigated in order to develop deception detection models that can be generalized across different contexts.

To obtain a precise detection, multimodal deception detection (MMDD) came into being. Following [7, 8], MMDD is a typical subtle visual computing task, aiming to detect imperceptible and deceptive clues from audio-visual scenarios. But general performance of cross-domain deception detection is unsatisfactory because it is challenging to reduce the domain gap between each dataset. In response to this, we propose the first Multimodal deception detection (MMDD) challenge - **SVC 2025**¹, aiming to bring together researchers and developers to advance the field of multimodal learning by detecting deception through the integration of multiple modalities such as audio, video, and text. The competition encourages innovation in building robust AI models that can accurately identify deceptive behaviors by leveraging various features from these modalities. Participants are required to submit their developed model, checkpoints and well-explained source code, accompanied by a paper describing their proposed methodologies and

the achieved results. Only contributions that meet the predetermined requirements, terms and conditions are eligible for participation. The organisers do not engage in active participation themselves, but instead undertake a re-evaluation of the findings of the systems submitted to challenge. The ranking of the submitted models depends on three metrics: accuracy, error rate, and F1-score for ranking on the test dataset split, with accuracy being the primary metric.

In summary, the main contributions and novelties of this challenge are listed as follows:

- We first introduce a new benchmark for the cross-domain audio-visual deception detection challenge. We present a comprehensive benchmark that evaluates the generalization capacity of AI models using audio and visual modalities across multiple domains in the deception detection task. Consequently, the SVC 2025 challenge requires models not only to be able to detect in a single domain, but also across multiple domains following three distinct domain sampling strategies, *i.e.*, domain simultaneous, domain alternating, and domain-by-domain.
- We introduce a novel protocol that enhances the model's generalization capability by evaluating performance across multiple domains, rather than limiting to a single-domain setting as in prior work. This better reflects real-world deployment conditions, where models must adapt to unseen or shifting data distributions, and therefore encourages the development of more robust and widely applicable solutions.

2 Related Works

2.1 Deception Detection Approaches

The research on using behavioral cues for deception has gradually become active over the past few decades. Among the studied behavioral cues, verbal and nonverbal cues were preferred as humans may behave differently between lying and telling the truth. Traditional deception detection is often a contact-based method. It assesses whether someone is telling the truth or not by monitoring physiological responses like skin conductance and heart rate [10, 13, 20]. Study [4] has revealed that, in general, people who tell lies are less forthcoming and less convincing than those who tell the truth. Liars usually talk about fewer details and make fewer spontaneous corrections. They also sound less involved but more vocally tense. Through the study, the researchers statistically found that liars often press their lips, repeat words, raise their chins, and show less genuine smiles. The results show that some behavioral cues do potentially appear in deception and are even more pronounced when liars are more motivated to cheat.

2.2 Multimodal Deception Detection

Recent advances in deception detection have increasingly incorporated both verbal and non-verbal cues, leveraging multimodal data to enhance detection performance. Gogate et al. proposed a deep model that incorporated the audio cues with visual and text modalities to improve the accuracy of

¹<https://sites.google.com/view/svc-mm25>

deception prediction. Karimi et al. explored deceptive cues from RGB images and raw audio in an end-to-end manner. Wu et al. utilized several types of features, including micro-expression and IDT (Improved Dense Trajectory) features from RGB images, MFCC (Mel-frequency Cepstral Coefficients) features from the audio, and transcripts.

Despite significant progress, there still exists the challenge of cross-domain generalization in multimodal deception detection. Previous works have mainly focused on optimizing unimodal or fusion methods within a single domain, without considering the domain shift when the system is applied to different environments or populations. For example, models trained on controlled lab data may not generalize well to real-world settings, where the recording conditions and communication styles may introduce significant variability.

In this challenge, we aim to follow the benchmark [8] for cross-domain generalization performance on the widely used audio-visual deception detection datasets, which is crucial for evaluating and improving the robustness of deception detection models in different scenarios. Establishing such a challenge will provide clearer comparisons, highlight model weaknesses, and guide the development of systems across different domains.

3 Challenge Corpora

The competition employed three datasets as training datasets: Real-life Deception Detection (Real-life Trial) [15, 16], Bag-of-Lies [9], and the Miami University Deception Detection Database (MU3D) [14], as shown in table 1. All participants must sign an agreement before accessing the datasets on their original platforms. Competition organizers will not provide raw data directly to participants. Instead, extracted OpenFace features, affect features from pretrained models, and Mel spectrograms (generated using PyTorch) are provided. These features do not contain any identifiable information.

The Real-life Trial Deception Dataset contains 121 video clips from real courtroom proceedings, averaging 28 seconds each. These clips include famous cases like the Jodi Arias trial, exoneration testimonies from "The Innocence Project," and defendant statements from crime-related TV episodes, featuring statements by defendants or witnesses. With 61 clips labeled deceptive and 60 truthful based on trial outcomes (guilty verdicts, not-guilty verdicts, and exonerations), the dataset covers 21 female and 35 male speakers aged 16–60. It also provides manually verified crowdsourced transcripts (8,055 words, including fillers/repetitions) and annotations of 9 non-verbal gesture categories (e.g., facial expressions, hand movements) using the MUMIN coding scheme.

The Bag-of-Lies dataset contains 325 annotated recordings from 35 unique subjects, including 162 deceptive and 163 truthful samples. It integrates four modalities: video capturing facial and body expressions via smartphone, audio of speech descriptions, gaze tracking data with fixation points and pupil metrics collected using Gazepoint GP3, and EEG signals from a 13-channel Emotiv EPOC+ headset sampled at 128 Hz. During collection, participants freely described 6-10 distinct images,

choosing spontaneously to lie or tell the truth per image. Recording durations ranged from 3.5 to 42 seconds.

The Miami University Deception Detection Database (MU3D) consists of 320 videos featuring 80 distinct participants of different races. Each participant generated four videos: a positive truth describing someone they genuinely liked, a negative truth describing someone they genuinely disliked, a positive lie falsely portraying a disliked person as liked, and a negative lie falsely portraying a liked person as disliked. The videos were collected in a laboratory setting where participants responded to standardized prompts for 45 seconds.

We use the Box of Lies game show dataset [19] for Stage 1 evaluation. This dataset contains 1,049 annotated utterances extracted from 25 publicly available video clips of The Tonight Show Starring Jimmy Fallon. The total video footage spans 2 hours and 24 minutes. It documents interactions between host Jimmy Fallon and 26 guests, including 6 males and 20 females, capturing both truthful and deceptive behaviors. Data collection involved extracting YouTube videos, segmenting conversational turns via ELAN software, and annotating multimodal behaviors such as facial expressions, head movements, and gaze according to the MUMIN coding scheme. Verbal content was transcribed via Amazon Mechanical Turk and manually verified for accuracy. Each utterance carries a veracity label, including 862 deceptive and 187 truthful instances. Linguistic features were extracted from transcripts, and non-verbal behaviors (e.g., smile frequency) were quantified as temporal percentages.

4 Evaluation Metrics

This competition employs three primary evaluation metrics: accuracy, error rate, and F1-score. Among these, **accuracy serves as the principal ranking criterion** for participant submissions. All metrics are computed based on binary classification outcomes where:

- Truthful samples are labeled as 1
- Deceptive samples are labeled as 0

- (1) **Accuracy:** Proportion of correctly classified samples

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

where:

- TP : True positives (correctly predicted deceptive)
- TN : True negatives (correctly predicted truthful)
- FP : False positives (truthful misclassified as deceptive)
- FN : False negatives (deceptive misclassified as truthful)

- (2) **Error Rate:** Complement of accuracy representing misclassification frequency

$$\text{Error Rate} = 1 - \text{Accuracy} = \frac{FP + FN}{TP + TN + FP + FN}$$

- (3) **F1-score:** Harmonic mean of precision and recall

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Dataset	#Subjects	Samples				Scenario
		Total	Deceptive	Truthful	Deception/Truth Ratio	
Real Life Trials[15]	56	121	61	60	1.02	Court-room
Bag of Lies[9]	35	325	162	163	0.99	Lab
MU3D[14]	80	320	160	160	1	Lab
Box of Lies[19]	26	1049	862	187	4.61	Gameshows

Table 1: Multimodal deception detection datasets in challenge

Predictions are generated as continuous scores $\in [0, 1]$ representing the probability of a sample being deceptive. For metric computation, these scores are thresholded at 0.5:

$$\text{Predicted class} = \begin{cases} 0 \text{ (deceptive)} & \text{if score} \geq 0.5 \\ 1 \text{ (truthful)} & \text{if score} < 0.5 \end{cases}$$

The evaluation process requires participants to submit prediction files containing sample paths and corresponding scores, formatted as:

SJ_BOL_EP3_lie_4 0.14431

5 Baseline model

We employ a cross-domain audio-visual deception detection method as the baseline model [8], extracting face frame features via ResNet18, acquiring behavioral features (AUs, gaze, and affect) using OpenFace and EmotionNet, and extracting Mel spectrograms for audio via OpenSmile or processing raw waveforms with Wave2Vec. Features from each modality are encoded and fused through modules like linear layers or Transformer before being fed into classifiers.

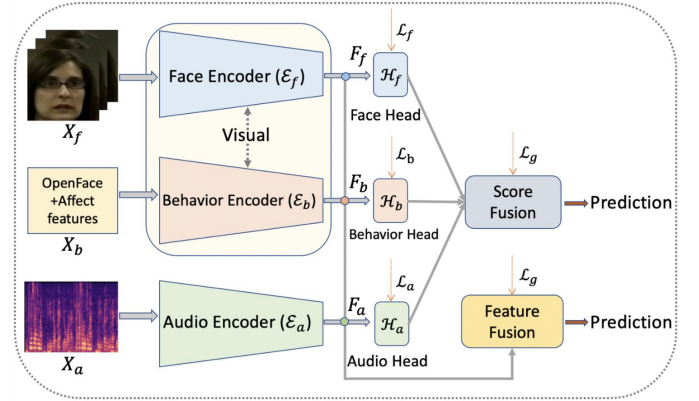
Cross-domain generalization strategies include single-to-single and multi-to-single scenarios. The latter introduces three sampling strategies:

- Domain-Simultaneous mixes samples from multiple source domains per batch to learn domain-invariant features;
- Domain-Alternating samples from one source domain batch by batch to capture domain specifics;
- Domain-by-Domain trains on source domains sequentially but may overfit to single-domain features.

The model architecture is depicted in Fig 2.

The MM-IDGM algorithm is proposed to align cross-domain gradient directions by maximizing gradient inner products between modality encoders. It dynamically adjusts learning rates based on the "Fish" algorithm and optimizes with previous unimodal losses to enhance multi-to-single generalization.

The Attention-Mixer fusion method combines MLP-Mixer and self-attention layers. It uses unimodal MLP layers to learn single-modality feature dynamics, self-attention layers to explore intra-feature associations, and cross-modal MLP layers to capture inter-modal interactions. With layer normalization and multi-head self-attention, six attention mixer layers are stacked: input features are projected, processed through multiple layers, and mean-pooled to output, enhancing intra- and inter-modal interactions for improved fusion performance.

**Figure 2: baseline model**

6 Participation

A total of 21 teams submitted their results by the end of the workshop competition ². The results of phase 2 are shown in Table. 2. We will introduce the approaches of some teams in the following part.

#	Team	ACC	F1	ERR
1	Glenn_xxy	62.437396	43.890274	37.562604
2	BigHandsome	60.434057	56.987296	39.565943
3	aim_whu	58.931553	45.333333	41.068447

Table 2: Challenge Results. ■ indicates the best result; ■ indicates the second-best result. Note that only the top-3 teams are shown here.

6.1 Team Glenn_xxy

Team Glenn_xxy proposed LCUNet, which employs three modality-specific branches (ResNet for audio and fusion, MLP for video) to extract initial features and logits from input images. Instead of simple 1x1 convolution for alignment, it introduces modality-specific projection networks to map these features into a unified space. These projection networks down-sample the feature dimensions to half the original size and then upsample back to the original scale, maintaining spatial

²<https://codalab.lisn.upsaclay.fr/competitions/22162#results>

resolution. The unified features are concatenated and processed by a deeper classifier to produce a fused logit. The final prediction combines the individual modality logits and the fused logit, leveraging both single-modality and fused discriminative information. The framework of proposed LCU-Net is shown in Fig 3.

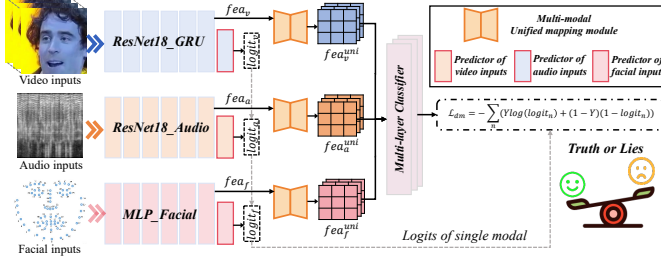


Figure 3: Network framework of Team Glenn_xxy

6.2 Team BigHandsome

Team BigHandsome, noting the significant domain gap across different datasets, employed a multi-loss framework to align training and testing features. Specifically, datasets are processed domain-by-domain. Multiple domain generalization loss functions are utilized simultaneously: CORAL loss for correlation alignment, Maximum Mean Discrepancy (MMD) loss for density divergence, Entropy Maximization (ATM) loss to confuse the domain discriminator, and Adversarial loss (based on CDAN+E principles) combined with a Gradient Reversal Layer for adversarial domain adaptation during backpropagation. The model schematic is presented in Fig 4.

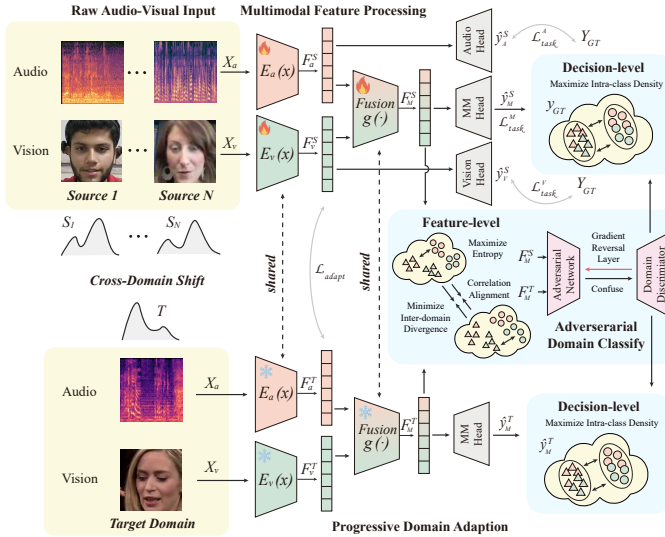


Figure 4: Network framework of Team BigHandsome

6.3 Team aim_who

Recognizing that ViT's global attention mechanism can more accurately capture fleeting and widespread cues in deceptive behaviors, Team aim_who selected Vision Transformer (ViT) as the facial feature extractor, replacing the baseline ResNet. Key training improvements include: Merging heterogeneous deception detection datasets to enhance sample diversity and generalizability; Adopting a Cosine Annealing LR scheduler (replacing StepLR) to achieve smooth learning rate decay and stable convergence; Implementing a differential learning rate strategy—applying a smaller learning rate to the pre-trained ViT backbone to preserve its knowledge, while using higher rates for other components. The model schematic is illustrated in Fig 5.

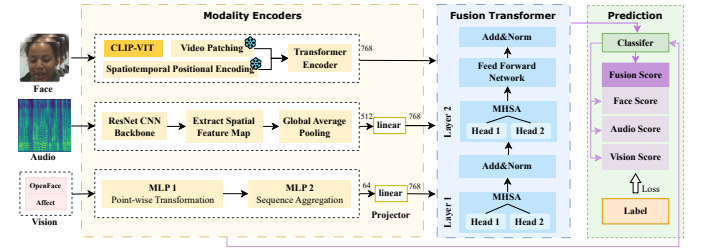


Figure 5: Network framework of Team aim_who

7 Discussion and Future Prospect

This challenge marks the first multimodal deception detection competition, serving as a valuable benchmark for the research community, fostering the development and application of multimodal data integration and fusion methods. Participants investigated visual, acoustic, and behavioral features etc, highlighting the potential of combining complementary multimodal information for superior deception detection. One notable strength of this challenge was its emphasis on intricate, real-world scenarios, leveraging diverse modalities and multiple datasets. This encourages innovative fusion strategies and the development of practical deception detection methods. However, one limitation was the short timeframe, which constrained participants from fully leveraging the raw multimodal data for deeper methodological exploration, particularly in areas such as temporal dynamics and end-to-end learning. Future challenge iterations could use longer timelines and bigger datasets for better models. Moreover, integrating large-scale pretrained models (e.g., multimodal foundation models) and introducing explainability tools would not only enhance predictive performance but also improve interpretability and trustworthiness of deception detection methods. These directions open exciting opportunities for advancing the field both in research and application.

8 Conclusion

As the competition demonstrates, we can see that multimodal deception detection still has great potential for development.

We hope that both the new challenge and baseline code, as well as the methodologies and outcomes of the participating teams, will serve as a helpful stepping stone for researchers who are interested in multimodal deception detection, leading to real-world applications.

Acknowledgments

References

- [1] Mohamed Abouelenien, Verónica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo. 2016. Detecting deceptive behavior via integration of discriminative features from multiple modalities. *IEEE Transactions on Information Forensics and Security* 12, 5 (2016), 1042–1055.
- [2] Charles F Bond Jr and Bella M DePaulo. 2006. Accuracy of deception judgments. *Personality and social psychology Review* 10, 3 (2006), 214–234.
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [4] Bella M DePaulo, James J Lindsay, Brian E Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological bulletin* 129, 1 (2003), 74.
- [5] Fletcher H Glancy and Surya B Yadav. 2011. A computational model for financial reporting fraud detection. *Decision support systems* 50, 3 (2011), 595–601.
- [6] Mandar Gogate, Ahsan Adeel, and Amir Hussain. 2017. Deep learning driven multimodal fusion for automated deception detection. In *2017 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 1–6.
- [7] Xiaobao Guo, Nithish Muthuchamy Selvaraj, Zitong Yu, Adams Wai-Kin Kong, Bingquan Shen, and Alex Kot. 2023. Audio-visual deception detection: Dolos dataset and parameter-efficient crossmodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22135–22145.
- [8] Xiaobao Guo, Zitong Yu, Nithish Muthuchamy Selvaraj, Bingquan Shen, Adams Wai-Kin Kong, and Alex C Kot. 2024. Benchmarking Cross-Domain Audio-Visual Deception Detection. *arXiv preprint arXiv:2405.06995* (2024).
- [9] Viresh Gupta, Mohit Agarwal, Manik Arora, Tanmoy Chakraborty, Richa Singh, and Mayank Vatsa. 2019. Bag-of-lies: A multimodal dataset for deception detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 0–0.
- [10] Hamza Javaid, Anika Dilawari, Usman Ghani Khan, and Bilal Wajid. 2022. EEG guided multimodal lie detection with audio-visual cues. In *2022 2nd International conference on artificial intelligence (ICAI)*. IEEE, 71–78.
- [11] Hossein Joudaki, Arash Rashidian, Behrouz Minaei-Bidgoli, Mahmood Mahmoodi, Bijan Geraili, Mahdi Nasiri, and Mohammad Arab. 2014. Using data mining to detect health care fraud and abuse: a review of literature. *Global journal of health science* 7, 1 (2014), 194.
- [12] Hamid Karimi, Jiliang Tang, and Yanen Li. 2018. Toward end-to-end deception detection in videos. In *2018 IEEE international conference on big data (Big Data)*. IEEE, 1278–1283.
- [13] Panfeng Li, Mohamed Abouelenien, Rada Mihalcea, Zhicheng Ding, Qikai Yang, and Yiming Zhou. 2024. Deception detection from linguistic and physiological data streams using bimodal convolutional neural networks. In *2024 5th International Conference on Information Science, Parallel and Distributed Systems (ISPDS)*. IEEE, 263–267.
- [14] E Paige Lloyd, Jason C Deska, Kurt Hugenberg, Allen R McConnell, Brandon T Humphrey, and Jonathan W Kunstman. 2019. Miami University deception detection database. *Behavior research methods* 51 (2019), 429–439.
- [15] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2015. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 59–66.
- [16] M Umut Şen, Veronica Perez-Rosas, Berrin Yanikoglu, Mohamed Abouelenien, Mihai Burzo, and Rada Mihalcea. 2020. Multimodal deception detection using real-life trial data. *IEEE Transactions on Affective Computing* 13, 1 (2020), 306–319.
- [17] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484–489.
- [18] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharmashan Kumaran, Thore Graepel, et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362, 6419 (2018), 1140–1144.
- [19] Felix Soldner, Verónica Pérez-Rosas, and Rada 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, Volume 1 (Long and Short Papers)*. 1768–1777.
- [20] John Synnott, David Dietzel, and Maria Ioannou. 2015. A review of the polygraph: history, methodology and current status. *Crime Psychology Review* 1, 1 (2015), 59–83.
- [21] Zhe Wu, Bharat Singh, Larry Davis, and V Subrahmanian. 2018. Deception detection in videos. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.