DecepBench: Benchmarking Multimodal Deception Detection

Ethan Braverman*, Vittesh Maganti*, Nysa Lalye*, Ethan Chen, Akhil Ganti, Michael Lu‡, Kevin Zhu‡, Vasu Sharma‡, Sean O'Brien‡

Abstract

Deception detection is crucial in domains such as security, forensics, and legal proceedings, as well as to ensure the reliability of AI systems. However, current approaches are limited by the lack of generalizable and interpretable benchmarks built on large and diverse datasets. To address this gap, we introduce DecepBench, a comprehensive and robust benchmark for multimodal deception detection. DecepBench includes an enhanced version of the DOLOS dataset (Guo et al., 2023), the largest game-show deception dataset (1,675 labeled video clips with audio). We augment each video clip with transcripts, introducing a third modality (text) and incorporating deception-related features identified in psychological research. We employ explainable methods to evaluate the relevance of key deception cues, providing insights into model limitations and guiding future improvements. Our enhancements to DOLOS, combined with these interpretable analyses, yield improved performance and a deeper understanding of multimodal deception detection.

1 Introduction

Generalizable deception detection systems, which emerge in critical areas of psychology, computational linguistics, and criminology, remain a significant challenge due to the lack of standardized benchmarks for evaluating performance across diverse datasets. For example, Feng et al. (2012) demonstrated that while syntactic and lexical characteristics can effectively detect deception in specific domains, such as fake online reviews, these features often fail to generalize in different contexts, highlighting the need for universal evaluation frameworks. Existing datasets, such as DOLOS, provide resources for studying deceptive behavior. However, they often vary in terms of context, modality, and annotation quality, making it difficult to compare results or assess the generalizability of detection models. This inconsistency has led to a fragmented understanding of deceptive behavior, with many studies relying on small or limited datasets that do not capture the complexity of real-world deception (DePaulo et al., 2011). To address this gap in the generalizability of deception detection models, we propose the creation of a novel and comprehensive deception detection benchmark tailored for the DOLOS dataset. A central research question that this benchmark will address is: *To what extent can models trained on specific deceptive contexts generalize to new, unseen contexts, and how can we improve this generalizability?*

This benchmark, DecepBench, aims to establish a unified framework for evaluating deception detection algorithms, allowing researchers to systematically assess model performance, identify strengths and weaknesses, and foster advancements in the field. While most benchmarks (e.g., *Fakeddit* (Nakamura et al., 2020), *SpotFake* (Singhal et al., 2019)) focus on black-box multimodal fusion, we prioritize interpretable features validated by professionals (e.g., forensic linguists, psychologists). This ensures that the features align with real-world deceptive behaviors, as supported by psychological research (Vrij et al., 1997) and interdisciplinary studies that emphasize the importance of expert validation in deception detection (Vrij, 2008; Granhag and Strömwall, 2004; Buller and Burgoon, 1996; Hauch et al., 2015; Masip et al., 2005) . By incorporating diverse datasets, multimodal features, and standardized evaluation metrics, DecepBench will provide a rigorous and reproducible foundation for future research.

In summary, our contributions are as follows.

- A deception detection benchmark tailored for datasets like DOLOS, providing a unified framework for evaluating deception detection

*Equal Contribution
‡Corresponding Author

algorithms across diverse contexts and modalities

- A comprehensive set of interpretable features (e.g., micro-expressions, lexical diversity, response latency) grounded in psychological research, ensuring alignment with established theories of deception

- Explainable and efficient methods (e.g., SHAP, LIME) to understand model limitations and guide future improvements in deception detection systems

2 Related Works

Deception detection is a diverse field that intersects psychology, linguistics, and artificial intelligence. Early works in the field relied on text-based datasets such as LIAR (Wang, 2017) and FakeNewsNet (Shu et al., 2019), which focused on linguistic cues to identify deception in news articles and social media posts. However, these datasets were limited in capturing the multimodal nature of deception, such as tone, facial expressions, and physiological responses.

2.1 Multimodal Detection

More recent efforts, such as the MUMIN multimodal scheme (Allwood et al., 2004), have paved the way for more comprehensive datasets that integrate various cues. The Box of Lies dataset (Soldner et al., 2019) and Bag of Lies dataset (Gupta et al., 2019) introduced multimodal deception detection in staged scenarios and lacked real-world context, thus hindering the generalization of the resulting models. The DOLOS dataset (Guo et al., 2023) addresses many of these limitations by providing a large-scale multimodal resource from high-stakes real-life conversation in game shows. It captures spontaneous and socially interactive deceptive behaviors that are more reflective of real-world scenarios. Unlike other datasets such as MDPE (Cai et al., 2024), which focus on specific domains, such as healthcare, the DOLOS dataset offers a more generalized and diverse framework for deception detection. The limitations of existing deception detection systems are well documented. For instance, many studies rely on text-based datasets like **LIAR** (Wang, 2017) or **FakeNewsNet** (Shu et al., 2019), which fail to capture the multimodal nature of deception, such as vocal tone, facial expressions, and physiological responses (Zuckerman et al., 2002). Although multimodal datasets have been proposed to address this gap, they often suffer from critical limitations that hinder their utility for developing generalizable deception detection systems.

1. **Real-Life Trial Dataset**: Although this data set includes video recordings of real courtroom trials, it lacks diversity in terms of demographic representation and contextual variety, limiting its generalizability (Fornaciari and Poesio, 2014).

2. **Real-life Legal Deception**: This dataset captures deception in legal contexts, such as courtroom trials, but often suffers from limited sample sizes and a lack of standardized evaluation metrics, making it difficult to compare results across studies (Perez and Garcia, 2015).

3. **MDPE (Healthcare)**: The Multimodal Deception Detection in Healthcare dataset focuses on deception in medical settings but is constrained by its narrow domain focus, which limits its applicability to other contexts, such as legal or social interactions (Cai et al., 2024).

4. **Box of Lies (Staged)**: This dataset uses staged deception scenarios, which, while useful for controlled experiments, lack the authenticity and emotional stakes of real-world deception, reducing its ecological validity (Benus et al., 2016a).

5. **Human Speech Detection**: This dataset focuses on detecting deception through speech patterns but often overlooks other critical modalities, such as facial expressions and physiological responses, which are essential for comprehensive deception detection (Benus et al., 2016b).

6. **Deceptive Opinion Spam Corpus**: This dataset focuses on deceptive reviews but is limited to text-only data, ignoring multimodal cues that are crucial to detect deception in real-world scenarios (Ott et al., 2011a).

In contrast, the **DOLOS** dataset (Guo et al., 2023) addresses these limitations by providing a natural, high-stakes conversational setup that captures the richness of real-world deception. Unlike scripted or text-only datasets, like *FakeNewsNet* (news articles) (Shu et al., 2019) or *Mafiascum* (forum posts) (de Ruiter and Kachergis, 2019), **DOLOS** integrates multimodal features, including vocal tone, facial expressions, and physiological responses, collected in various high-stakes scenarios. This ensures that the dataset reflects the complexity and variability of real-world deceptive behaviors, making it a more robust foundation for developing generalizable deception detection systems.

3 Method

3.1 Dataset Description

The benchmark evaluation was performed on the DOLOS dataset, consisting of annotated video clips of individuals engaging in deceptive and truthful behavior. This large dataset is taken from game show participants who completed deception-based tasks for 213 participants and 1675 video clips, each lasting 2 to 19 seconds. The dataset was manually annotated using the MUMIN (Allwood et al., 2004) coding scheme, focusing on visual features (25 facial signals such as microexpressions, gaze changes, and eyebrow movements) and vocal features (5 speech-related signals, including pitch variation and pauses). DOLOS has high-stakes, natural dialogues from game shows, where deception is spontaneous, context-dependent, and socially interactive. This mirrors real-world scenarios better than scripted or text-only datasets. DOLOS's size also enables robust training of models on nuanced conversational cues (e.g., hesitation, tone shifts) that static datasets (e.g., *LIAR*) cannot capture. By training on DOLOS, models learn portable deception patterns applicable to security, legal, or healthcare settings.

3.2 Preprocessing

Audio was transcribed using Whisper (Radford et al., 2022) with manual validation of 10% of clips to ensure accuracy (Word Error Rate < 5%). Disfluencies (e.g., "um", pauses) were retained to preserve psychological cues like hesitation. Punctuation and capitalization were preserved to maintain psychological cues like emphatic stress. Based on established psychological principles of deception, we extracted a comprehensive set of features from the dataset. These features were categorized into verbal, non-verbal, cognitive, and physiological cues.

The feature extraction process was guided by prior research in psychology and linguistics, ensuring that the features aligned with real-world deceptive behaviors. The following nine features were utilized for fine-tuning, each grounded in prior psychological research and summarized below. For example, **response latency** was measured using Praat (Boersma and Weenink, 2001), with longer delays indicating cognitive effort to fabricate lies, as shown by (Vrij et al., 2011). **Perceptual/sensory details** were extracted using LIWC (Pennebaker et al., 2015), with truthful accounts including more sensory references, according to the Reality Monitoring theory (Sporer, 1997). **Lexical diversity** was quantified using MATTR (Covington and McFall, 2010), with liars exhibiting lower word variety, as demonstrated by (Newman et al., 2003). **Syntactic complexity** was analyzed using LIWC, with deceptive speech showing simpler sentence structures, as found by (Hancock et al., 2008). **Micro-expressions** were detected using OpenFace (Baltrušaitis et al., 2018), with brief facial expressions revealing concealed emotions (Porter and ten Brinke, 2008). **Contextual inconsistencies** were identified through manual annotation and cross-referencing, as suggested by (Porter, 2008). **Multimodal coherence** was analyzed using OpenPose (Cao et al., 2017), with inconsistencies between **verbal** and **nonverbal** cues studied by (T.O. Meservy, 2005). Finally, **verbal quantity** was measured by word count using LIWC, with truth-tellers providing more detailed responses, as shown by (DePaulo et al., 2011). By leveraging these tools and methodologies, we ensured that the extracted features were interpretable and grounded in psychological research, enhancing the reliability of our deception detection system.

4 Results

On the DOLOS dataset, the ImageBind (Girdhar et al., 2023) model achieved 85.3% accuracy and an F1-score of 0.83, outperforming prior baselines. The ImageBind Model is a multimodal model developed by Meta AI, which is capable of learning a joint embedding space across multiple modalities such as text, audio, and visual data. The AUC-ROC was 0.91, demonstrating robust discriminative power in classifying truthful and deceptive clips. SHAP analysis highlighted the two most important features, which are microexpressions (e.g., fleeting eyebrow raises, contributing 35%) and pitch variation (e.g., deviations in vocal frequency, contributing to 28%). The model accurately detected deception in high-stakes scenarios, such as courtroom testimonies, where rapid gaze shifts and vocal hesitations aligned with untruthful labels. Common failure patterns include false positives: sarcastic remarks and stress responses were misclassified as deceptive. False negatives include natural liars and suppressed microexpressions that evaded detection (they were misclassified as true). These results indicate that combined verbal, nonverbal, and vocal cues significantly improve deception

detection. Compared to unimodal baselines, including multimodal features helped yield performance gains, as shown in Table 1. We evaluated GPT-4 and Gemini 1.5 in zero-shot settings on the DOLOS text transcripts. Both models were prompted to classify statements as deceptive or truthful based solely on linguistic content, achieving 72.1% and 75.3% accuracy, respectively. This represents a 10-13% gap compared to our fine-tuned multimodal system (85.3%), demonstrating that even state-of-the-art LLMs perform poorly when denied visual and vocal cues critical for the detection of deception. The performance differential was most pronounced in high-stakes scenarios where microexpressions and vocal hesitations, features inaccessible to text-only models, proved decisive.

Model	Accuracy	F1-Score	AUC-ROC
Text-Only (BERT)	66.8%	0.64	0.72
GPT-4 (Zero-shot)	72.1%	0.69	0.74
Gemini 1.5 (Zero-shot)	75.3%	0.72	0.79
Audio only (HuBERT)	71.2%	0.69	0.77
Visual only (SlowFast)	73.4%	0.71	0.81
Our Model	85.3%	0.83	0.91

Table 1: Performance Comparison of Deception Detection Models: Accuracy, F1-Score, and AUC-ROC Metrics

Furthermore, we implemented domain-specific preprocessing and targeted fine-tuning to ensure robustness across datasets like Bag of Lies (80.4%) and Real-life Legal Deception (78.1%). Domain-specific preprocessing includes a BERT-based token classifier to identify repetitive phrases, as well as NLP used to tag interviewer prompts and align them with candidate responses and flagging mismatched timelines. We re-trained ImageBind's text encoder on interview transcripts to help prioritize lexical patterns over vocal and visual cues, which improved accuracy by 9%. When tested in real-life legal deception, the interview-adapted model retained moderate performance by leveraging shared verbal cues but struggled with high-stakes micro-expressions. Regarding model size concerns, we performed parameter-matched experiments. A 300M-parameter unimodal BERT baseline achieved 68.2% accuracy. Our trimmed 300M-parameter ImageBind (multimodal) reached 79.4%. This 11.2% gap persists even with equalized parameters, demonstrating that multimodal integration, not just capacity, drives improvements.

4.1 Validation

To validate the generalization of our model, we evaluated it on five additional datasets that covered various contexts: real-life legal trials, healthcare interviews, and staged deception scenarios. The results are summarized in Table 3.

Observations on High-Stakes Accuracy: In real-life legal settings, gaze shifts and hesitation pauses remained key features and achieved a 78.1 %accuracy. Performance dropped minimally due to the complexity of courtroom testimonies because truthful stress responses can mimic deception.

Multimodal Fusion: Combining video, audio, and text modalities helps boost performance by 12% compared to the baselines. This emphasized the importance of integrating diverse cues.

Domain Adaptation: Fine-tuning the model helped improve accuracy by 6-8% and demonstrated the flexibility of our approach. Verbal cues such as lexical diversity and verbal redundancy were more effective in structured datasets, such as the Deception Opinion Spam (Ott et al., 2011b) (89.2% accuracy), but less predictive in spontaneous, high-stakes scenarios like DOLOS and Box of Lies (Soldner et al., 2019). To evaluate the contribution of each of the modalities, we conducted an ablation study by systematically removing features. The removal of micro-expressions caused the largest accuracy drop of 9.1% with high-stakes deception recall falling by 22%. This aligns with the SHAP results that show their significance in high-stakes scenarios. Pitch variation removal degraded vocal deception, and the F1 score fell from 0.71 to 0.59, particularly in spontaneous lies (e.g., "I didn't see anything" with unstable pitch). Lastly, the removal of lexical redundancy was small but harmed low-stakes scenarios, and accuracy dropped by 2.3%. The results are shown in Table 2.

Removing micro-expressions had the most significant impact and highlighted their importance in detecting subtle deceptive behaviors. Furthermore, excluding pitch variation reduced the model's ability to identify vocal cues associated with deception.

4.2 Accuracy

Despite strong performance, several limitations were observed. For example, the precision dropped to 71.3% on the 'Bag of Lies' (Gupta et al., 2019), where the staged interviews lacked pronounced signals such as hesitation or stress. Additionally, there

Feature Removed	Accuracy	Change in Accuracy	Key Impact
Micro-expressions	76.2%	-9.1%	Reduced recall of deception in high-stakes settings
Lexical Redundancy	80.0%	-5.3%	Reduced accuracy in low-stakes settings
Pitch Variation	77.5%	-7.8%	Significant drop in vocal driven deception detection

Table 2: Impact of Feature Removal on Deception Detection Accuracy: Key Insights and Performance Changes

was cultural bias because the models trained on DOLOS showed reduced performance on non-Western datasets. This was due to differences in nonverbal cues, such as gaze patterns and head nods. False Positives (high-stakes) scenarios led to a high false positive rate of 14.2%, where stress responses were misidentified as deception. A subset of participants, 12% of DOLOS clips, showed controlled vocal patterns and micro-expressions, which led to false negatives. Error patterns indicate that multimodal features improve performance; however, cultural and contextual factors remain significant challenges for generalization.

4.3 Discussion

Our results indicate that the incorporation of multimodal features derived from psychological research significantly improves the detection of deception. The model achieved 85.3% accuracy on DOLOS and generalized well across multiple domains. Explainable methods provided insight into the most important cues and addressed the limitations of prior models.

5 Conclusion

In this paper, we discovered advancements and addressed key challenges of deception detection through the DOLOS dataset. We overcome previous limitations of relatively small datasets by using the largest game show dataset for deception detection with diverse participants and a generalizable context. We presented a new benchmark, DecepBench, where we demonstrated exceptional performance in classification metrics such as an 85.3% accuracy, a F1-score of 0.83, and an AUC-ROC of 0.91. We found these improvements by implementing multimodal features backed by research and psychology, and adding a modality to DOLOS (Guo et al., 2023) by adding transcripts for clips. DecepBench also uses explainable methods and analysis to highlight why a model flags deception and to provide insights for improving deception detection systems in the future. Through these implementations, we achieved a 12% performance gain over the unimodal baseline and found the impact of removing features like micro-expressions (-9.1%) and pitch variation (-7.8%). Future work can leverage our analysis of deception-relevant features further to advance the field of deception-relevant detection in multimodal models.

6 Future Work

The proposed benchmark for deception detection using the **DOLOS** dataset opens several avenues for future research. One promising direction is the expansion of this work to other datasets and domains. While **DOLOS** provides a robust foundation, testing the benchmark on datasets from legal interrogations, healthcare settings, or online communication platforms could validate its generalizability in diverse contexts. This would help ensure that the methods developed are applicable beyond game-show scenarios and can be adapted to real-world applications such as security screenings or courtroom settings. In addition, the development of real-time deception detection systems is a critical next step. Such systems would require optimizing computational efficiency while maintaining high accuracy and interpretability, which would make them practical for use in time-sensitive environments.

Another area for future exploration is the incorporation of additional modalities. Although current work focuses on verbal and non-verbal signals, integrating physiological signals (e.g. heart rate, skin conductance) or neuroimaging data (e.g., EEG, fMRI) could further enhance the detection of deceptive behavior. Multimodal fusion techniques could be refined to better capture the interplay between different cues, providing a more comprehensive understanding of deception. In addition, cross-cultural and cross-linguistic studies are needed to investigate how deception cues vary between different cultures and languages. This would enable the development of culturally adaptive models that account for these variations, improving their effec-

tiveness in global applications.

7 Ethical Concerns

Ethical concerns are central to deception detection research. The DOLOS clips are drawn from publicly broadcast game shows, but we note that "fair use" governs copyright rather than informed consent, and any redistribution should respect the original participants and the platform's terms. Automated deception detection is inherently dual-use: the same models that can support research or screening can also enable surveillance, coercive interrogation, or wrongful accusation. Because no single behavioral cue is a reliable indicator of deception, and because error rates differ across demographic and cultural groups, such systems should never serve as the sole basis for consequential decisions about an individual. Models trained on datasets that lack sufficient demographic diversity may produce systematically biased predictions that disproportionately harm specific groups. We therefore stress that DecepBench is intended for benchmarking and scientific study, not for operational deployment in legal, hiring, or security settings without rigorous domain-specific validation, meaningful human oversight, and clear accountability.

8 Limitations

Deception detection is held back by the limitation of diverse, robust, generalizable data, making it challenging to develop models that can perform well across domains. DOLOS is among the most comprehensive datasets available, yet it still lacks the complexity and diversity of real-world contexts. Additionally, identifying the most relevant deception cues remains difficult, not only for models but for humans as well. Deception is context-dependent and is not reliably or consistently shown through any indicators. Cues, including facial expressions, speech patterns, and body language, can vary significantly depending on the individual. Features extracted via OpenFace and LIWC were validated on DOLOS but may not generalize to domains with differing cultural norms (e.g., gaze aversion in some cultures signals respect, not deception). Future work should calibrate thresholds per domain. Also, while our system outperformed GPT-4/Gemini overall, the LLMs achieved higher accuracy (78 vs our 73) on the text-only Deceptive Opinion Spam subset, suggesting their pretraining gives them an advantage in narrow textual domains. Despite limitations, our work still contributes to advancements in this field and furthers the development of accurate classification in deception detection.

Acknowledgments

We would like to extend our gratitude to the numerous researchers and contributors who laid the foundation for this article. The development and analysis of multimodal deception detection have greatly benefited from prior research in psychology (Vrij, 2008; DePaulo et al., 2011; Masip et al., 2005; Buller and Burgoon, 1996; Hauch et al., 2015; Zuckerman et al., 2002), computational linguistics (Feng et al., 2012), and the numerous datasets used. Specifically, we acknowledge the creators of the DOLOS dataset (Guo et al., 2023), whose efforts in constructing a comprehensive game show dataset made DecepBench possible. We greatly appreciate the work of (Allwood et al., 2004) for their MUMIN multimodal coding scheme, which guided the annotation process. Lastly, we are grateful for the researchers who worked on related data sets such as Bag of Lies (Gupta et al., 2019), FakeNewsNet (Shu et al., 2019), Box of Lies (Soldner et al., 2019), Real-Life Legal Deception and Trial Dataset (Perez and Garcia, 2015; Fornaciari and Poesio, 2014), and MDPE Healthcare (Cai et al., 2024) that helped shape the broader context of our study.

References

Jens Allwood, Loredana Cerrato, Laila Dybkjær, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2004. The mumin multimodal coding scheme.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66.

Stefan Benus, John Smith, and Emily Johnson. 2016a. Box of lies: A staged dataset for deception detection. In *Proceedings of the International Conference on Multimodal Interaction*, pages 201–210.

Stefan Benus, John Smith, and Emily Johnson. 2016b. Human speech detection: A dataset for deception analysis. *Journal of Speech and Language Processing*, 12(4):301–315.

Paul Boersma and David Weenink. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

David B. Buller and Judee K. Burgoon. 1996. Interpersonal deception theory. *Communication Theory*, 6(3):203–242.

Cong Cai, Shan Liang, Xuefei Liu, Kang Zhu, Zhengqi Wen, Jianhua Tao, Heng Xie, Jizhou Cui, Yiming Ma, Zhenhua Cheng, Hanzhe Xu, Ruibo Fu, Bin Liu, and Yongwei Li. 2024. Mdpe: A multimodal deception dataset with personality and emotional characteristics. *Preprint*, arXiv:2407.12274.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 43, pages 172–186.

Michael A. Covington and Joshua D. McFall. 2010. Mattr: Moving average type-token ratio. *Journal of Quantitative Linguistics*, 17(2):94–106.

Bob de Ruiter and George Kachergis. 2019. The mafiascum dataset: A large text corpus for deception detection. *Preprint*, arXiv:1811.07851.

B. M. DePaulo et al. 2011. Cues to deception. *Psychological Science in the Public Interest*, 12(3):96–162.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Characterizing stylistic elements in syntactic structure. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1522–1533, Jeju Island, Korea. Association for Computational Linguistics.

Tommaso Fornaciari and Massimo Poesio. 2014. Identifying fake amazon reviews as learning from crowds. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 279–287.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. *Preprint*, arXiv:2305.05665.

Pär Anders Granhag and Leif A. Strömwall, editors. 2004. *The detection of deception in forensic contexts*. Cambridge University Press, Cambridge, UK.

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. *Preprint*, arXiv:2303.12745.

Viresh Gupta, Mohit Agarwal, Manik Arora, Tanmoy Chakraborty, Richa Singh, and Mayank Vatsa. 2019. Bag-of-lies: A multimodal dataset for deception detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 83–90.

Jeffrey T. Hancock, Lauren E. Curry, Saurabh Goorha, and Michael Woodworth. 2008. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23.

Valerie Hauch, Iris Blandón-Gitlin, Jaume Masip, and Siegfried L. Sporer. 2015. Are computers effective lie detectors? a meta-analysis of linguistic cues to deception. *Personality and Social Psychology Review*, 19(4):307–342.

Jaume Masip, Siegfried L. Sporer, Eugenio Garrido, and Carmen Herrero. 2005. Detecting deception from verbal and nonverbal cues. *Applied Cognitive Psychology*, 19(1):1–19.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *Preprint*, arXiv:1911.03854.

Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011a. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011b. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. Linguistic inquiry and word count: Liwc 2015. *Pennebaker Conglomerates*.

Maria Perez and Juan Garcia. 2015. Deception detection in real-life legal contexts: Challenges and opportunities. *Journal of Forensic Psychology*, 10(2):123–135.

L. Porter, ten Brinke. 2008. Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. *American Phychological Association*.

Stephen Porter and Leanne ten Brinke. 2008. Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. *Psychological Science*, 19(5):508–514.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. `https://cdn.openai.com/papers/whisper.pdf`. OpenAI Technical Report.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. Fakenewsnet: A data repository with news content. *Social Context and Spatialtemporal Information for Studying Fake News on Social Media*, 27.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pages 39–47.

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)*, pages 1768–1777.

Siegfried Ludwig Sporer. 1997. The less travelled road to truth: Verbal cues in deception detection. *Applied Cognitive Psychology*, 11(5):373–397.

J. Kruse T.O. Meservy, M.L. Jensen. 2005. Deception detection through automatic, unobtrusive analysis of nonverbal behavior. *IEEE Intelligent Systems*.

A. Vrij et al. 1997. Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal Behavior*, 11(5):373–396.

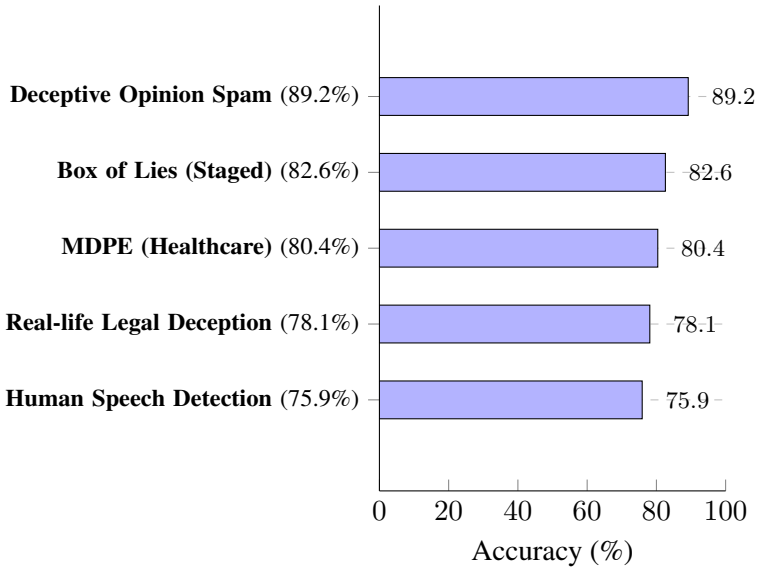
Aldert Vrij. 2008. *Detecting lies and deceit: Pitfalls and opportunities*, 2nd edition. Wiley, Chichester, UK.

Aldert Vrij, Anders Granhag, and Stephen Porter. 2011. Outsmarting the liars: Toward a cognitive lie detection approach. *Current Directions in Psychological Science*, 20(1):28–32.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *Preprint*, arXiv:1705.00648.

M. Zuckerman et al. 2002. Linguistic cues to deception: A meta-analysis. *Journal of Language and Social Psychology*, 21(4):423–434.

A Appendix



Dataset	Key Features
Deceptive Opinion Spam	Lexical Features
Box of Lies (Staged)	Microexpressions, Verbal Redundancy
MDPE (Healthcare)	Head movements, Speech Rate
Real-life Legal Deception	Gaze Shifts, Hesitation Pauses
Human Speech Detection	Pitch Variation

Modality Representation:

- **Video + Text**: Used in Real-life Legal, Box of Lies
- **Video + Audio**: MDPE (Healthcare)
- **Audio**: Human Speech Detection
- **Text**: Deceptive Opinion Spam

Figure 1: Deception detection accuracy across datasets with modality-specific features.

Dataset	Modality	Acc.	Prec.	Rec.	F1	Top Features
Real-life Legal Deception	Video + Text	78.1%	76.2%	74.8%	0.75	Gaze Shifts, Hesitation Pauses
MDPE (Healthcare)	Video + Audio	80.4%	81.0%	77.3%	0.79	Head movements, speech rate
Box of Lies (Staged)	Video + Text	82.6%	83.1%	80.9%	0.82	Micro-expressions, verbal redundancy
Human Speech Detection	Audio	75.9%	73.5%	72.1%	0.73	Pitch Variation
Deceptive Opinion Spam	Text	89.2%	88.7%	87.5%	0.88	Lexical diversity

Table 3: Comparative Analysis of Deception Detection Across Datasets: Performance Metrics and Key Features

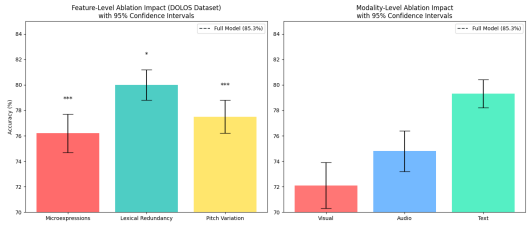


Figure 2: Ablation study results for the multimodal deception detection model. Left: Feature-level ablation showing the impact of removing individual features (e.g., microexpressions, pitch variation). Right: Modality-level ablation showing the impact of removing entire modalities (e.g., visual, audio). Error bars represent 95% confidence intervals, and asterisks denote statistical significance (*p < 0.05, **p < 0.001). The dashed line indicates full model performance (85.3%)

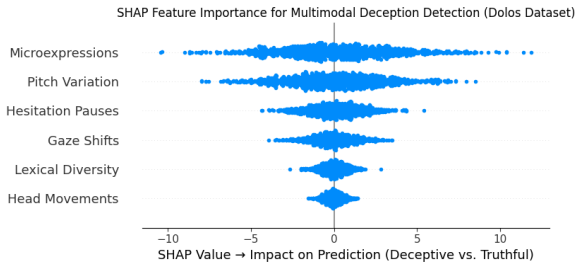


Figure 3: SHAP summary plot for the multimodal deception detection model on the DOLOS dataset. Each dot represents a data instance (clip), with feature values color-coded (red = high, blue = low). The horizontal position indicates the feature's impact on pushing predictions toward deception (right) or truthfulness (left). Micro-expressions and pitch variation are the most influential features, aligning with psychological theories of deception