Multimodal Deception Detection System

Doaa Yahia
Computer Science Department,
Faculty Of Computer and Information Sciences,
Ain Shams University,
Cairo Egypt
20201700243@cis.asu.ed.eg

Medhat Essam
Computer Science Department,
Faculty Of Computer and Information Sciences,
Ain Shams University,
Cairo Egypt
20201700789@cis.asu.edu.eg

Mena Ashraf
Computer Science Department,
Faculty Of Computer and Information Sciences,
Ain Shams University,
Cairo Egypt
20201700888@cis.asu.edu.eg

Mohamed Moawad
Computer Science Department,
Faculty Of Computer and Information Sciences,
Ain Shams University,
Cairo Egypt
20201700743@cis.asu.edu.eg

Reem Ayman
Computer Science Department,
Faculty Of Computer and Information Sciences,
Ain Shams University,
Cairo Egypt
20201700293@cis.asu.edu.eg

Rodayna Mohamed
Computer Science Department,
Faculty Of Computer and Information Sciences,
Ain Shams University,
Cairo Egypt
20201700284@cis.asu.edu.eg

Dr. Salsabil Amin,
Basic Science Department,
Faculty of Computer and Information Science,
Ain Shams University

T.A. Zeina Rayan,
Computer Science Department,
Faculty of Computer and Information Sciences,
Ain Shams University.

***Abstract*—** Deception is a common issue with significant consequences in areas such as police investigations, airport security, and courtroom trials. Human ability to detect deception without specialized tools is quite limited. Deceptive behavior often triggers multiple cues, including inconsistencies in linguistic choices, changes in tone, hesitation, nervous gestures, avoidance of eye contact, and incongruent facial expressions. Detecting these cues is complex, requiring professional training, which can be costly and time-consuming, and still may not guarantee accuracy.

Our goal was to create an artificial intelligence (AI) system that can recognize dishonesty using both verbal and visual clues, thereby offering a practical substitute for expert instruction. We used a multimodal strategy that included micro-expressions, textual linguistic traits, audio signals, and video data. We developed and tested our models from scratch by investigating and improving upon earlier standards. Our system used late fusion of visual, acoustic, and linguistic information to achieve 95.8% accuracy on the Real-life Deception Detection Dataset. This solution shows promise and considerable advancements.

1. Introduction

Deception involves actions intended to mislead, conceal the truth, or promote false beliefs. Detecting deception is critical in security-sensitive areas like police investigations and airport security. Traditional methods, such as polygraph tests [7], require skin-contact devices and human expertise, making them impractical and susceptible to human error and bias [6]. Moreover, offenders can employ countermeasures to deceive these methods.

Learning-based approaches using text [13], speech [8], and facial expressions [9] have been proposed to address these limitations. Research indicates that micro-expressions—brief, involuntary facial movements—are significant indicators of deception.

Modern systems have integrated multiple modalities, combining text, audio, and visual data to enhance detection accuracy [5], [10], [11].

Our research aims to develop a multimodal deception detection system leveraging video, audio, linguistic features from text, and micro-expressions. This comprehensive approach addresses the limitations of traditional methods and offers a robust solution for detecting deceptive behavior in critical settings.

Currently, law enforcement and airport security rely heavily on human judgment, which has a low accuracy rate of about 54% in detecting deception [1]. This highlights the need for more accurate detection systems. Human judgment can be influenced by biases and external factors, making it unreliable. Artificial intelligence can significantly improve deception detection by analyzing facial micro-expressions from pre-recorded videos, providing a more reliable and unbiased assessment.

2. Dataset

2.1 Dataset Overview

To evaluate our deception detection model, we utilized a real-life deception detection dataset as referenced in [3]. This dataset comprises 121 video clips of courtroom trials, with 61 clips classified as lies and the remaining 60 classified as truth.

Figure 1: Dataset screenshots

The dataset is annotated with micro-expressions identified in each video. An accompanying Excel sheet includes columns for 39 possible micro-expressions (see Figure 2).

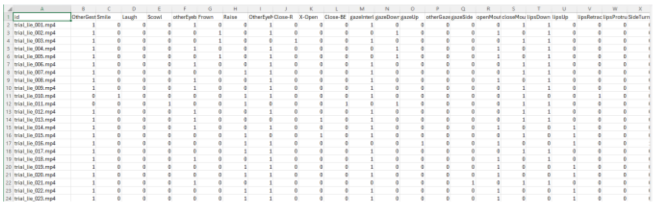

Figure 2: Micro expressions annotations

However, we also extracted additional micro expressions beyond those provided in the dataset.

Additionally, the dataset includes transcriptions for each video (see Figure 3).

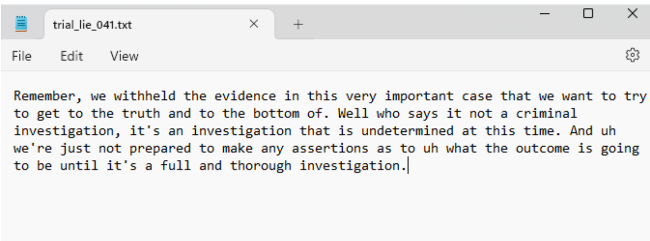

Figure 3: Transcriptions for video

Some videos in the dataset had issues, such as multiple subjects in the frame simultaneously and dialogues between two people. These problems were addressed by manually cropping and clipping the videos, as no automated solutions were available. Additionally, a few videos were severely corrupted and unsuitable for training, so they were omitted from our testing.

The dataset also has bias issues. The same subjects often reappear in multiple videos, exhibiting consistent behaviour, which can introduce bias. Moreover, there is a disproportionate gender representation in deceptive videos, which can further skew results if the data is split randomly. To mitigate this, we manually split the dataset during our experiments to ensure that the training and test sets did not have overlapping subjects.

3. METHODOLOGY

3.1 Preprocessing and feature extraction

Video Frames Preprocessing

In our approach, we utilized each video sample in its entirety rather than segmenting it into multiple parts, as done in other approaches [9], [19]. Segmentation can introduce inaccuracies since the exact moment of deception within a video segment is uncertain. To maintain consistency, we downsampled the videos to achieve a uniform sequence length, treating each testimony as a single input for our models during both training and testing.

Micro-Expressions Extraction

Although the dataset includes manually annotated expressions for deception detection, we opted to use the Py-Feat library [22] to automate the annotation process. Py-Feat extracts various signals and facial features from videos, focusing on Action Units (AUs) and emotions critical for deception detection. Features were extracted every 30 frames to balance computational efficiency and detail, and the resulting records were combined into a single mean feature vector. For our models, we selected Retina Face for face detection, Mobilefacenet for landmarks, XGB for Action Units, and Resmasknet for emotions, ensuring a comprehensive and automated feature extraction process.

Text Preprocessing

The transcribed text underwent preprocessing within an NLP pipeline, which involved removing punctuation for standardization, converting words to lowercase for consistency, splitting the text into individual words, filtering out stop words to enhance text quality, and reducing words to their base forms. Finally, the cleaned text was transformed using a vectorizer like TF-IDF or GloVe to enable numerical representation suitable for machine learning and deep learning models.

Audio Preprocessing

In this audio modality, we extracted the Mel-frequency cepstral coefficients (MFCC) from audio signals to reveal deceptive patterns, as supported by previous research [8]. The preprocessing pipeline involved extracting audio from the video, performing Short-Time Fourier Transform (STFT) noise reduction to clean the audio, and extracting MFCC features at specified framing windows with a frame length of 0.025 seconds and a hop length of 0.01 seconds at a target sampling rate of 44,100 Hz. To address the varying durations of videos and achieve a consistent input shape for our sequential models, we downsampled the MFCC segments.

3.2 Models Implementation

In our proposed system, multiple neural network and machine learning models analyze various input data types, including text, video, audio, and micro expression data. The final prediction from these models classifies something as either "Truthful" or "Deceptive" after combining their outputs using a late fusion technique (see Figure 4).

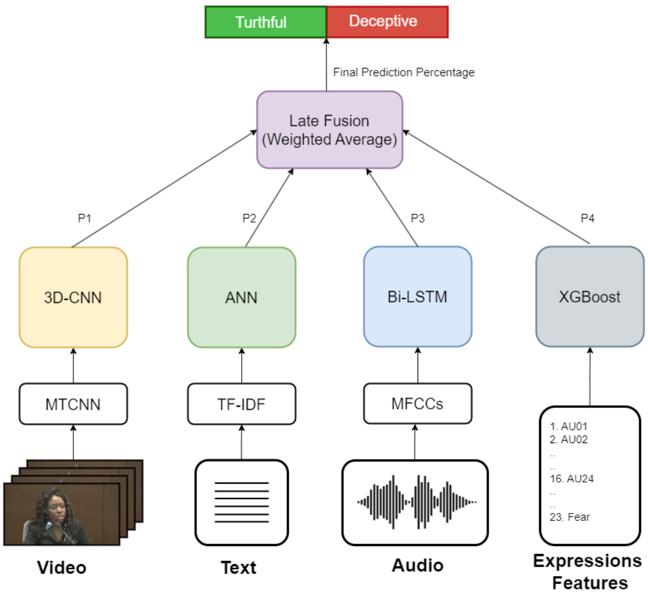

Figure 4 System Model Architecture

Video Frames Modality

Lie detection via video utilizes involuntary facial expressions, eye movements, and body gestures indicative of deception, with previous research [9], [10], [11], [19] demonstrating the effectiveness of neural networks based on CNN architectures in capturing these patterns. We experimented with two neural network architectures: CNN-LSTM and 3D CNN, both of which combine spatial and temporal feature extraction for final classification.

Micro Expressions Modality

After extracting the mean Action Units and emotions feature vector from each video sample using Py-Feat, we employed machine learning classifiers to identify deceptive cues in these micro expressions. Supported by prior research [3], [5] demonstrating their effectiveness in deception detection, we chose machine learning classifiers over deep learning neural networks due to the less complex nature of this data modality and the limited number of samples, which make neural networks prone to overfitting. We experimented with algorithms such as Support Vector Machine (SVM), Decision Tree, Random Forest, and Extreme Gradient Boosting (XGBoost).

Text Modality

We utilized vectorized textual data to detect deceptive patterns, employing various machine learning algorithms and simple deep learning models such as a basic Artificial Neural Network (ANN). Given the limited number of samples, more complex models like Recurrent Neural Networks (RNNs) and Transformers were unsuitable due to their propensity for overfitting or failure to learn patterns effectively. The machine learning algorithms employed included Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), and Extreme Gradient Boosting (XGBoost), while the ANN architecture consisted of two hidden layers and a final sigmoid layer for prediction.

Audio Modality

The extracted MFCCs were used to analyze subtle vocal cues such as pitch, tone, and speech pace. We employed Recurrent Neural Networks (RNNs) to detect deception by analyzing these MFCC sequences, experimenting with simple RNN-based architectures including GRUs, LSTMs, and Bidirectional LSTMs, followed by a final sigmoid layer for classification.

Data Fusion

Finally, data fusion creates more accurate and thorough results by combining information from several modalities.
It combines several forms of data, improving comprehension and judgment. In our approach, we experimented with late fusion.

3.3 Results

Table 1 presents a comprehensive summary of our deception detection system's performance, comparing unimodal and multimodal approaches.

These results underscore the superiority of multimodal approaches, particularly majority voting, in enhancing deception detection accuracy. The substantial improvement from 87.5% in the best unimodal method to 95.8% in the best multimodal approach highlights the synergistic effect of combining multiple data sources for this task (see Table 1).

Table 1: Results Summary

Unimodal Results				
Modality	Video	Expressions	Text	Audio
Model	3D-CNN	XGBoost	ANN	Bi-LSTM
Accuracy	87.5%	75%	79.1%	79.1%
Multimodal Results				
Late Fusion Technique	Weighted average	**Majority Voting**	Weighted average	Majority Voting
Features	A+V+T	**A+V+T**	A+V+T+E	A+V+T+E
Accuracy	91.6%	**95.8%**	87.5%	91.6%

Competitive analysis

Comparing our best results with some of the state-of-the-art multimodal approaches results, particularly the results in [5], [9], [10], [11], [20].

The table below highlights the methodologies used in previous works and their respective accuracy and how our approach compares to it.

Our system performs reasonably well in comparison to other approaches, however, the approach provided by Krishnamurthy, Gangeshwar, et al [11] performs a little bit better but that is expected since their system is semi-automatic (it relies on manual annotations for micro expressions) while our system is fully automated (see Table 2).

Table 2: Competitive analysis

Citation	Dataset	Methodology	Best Accuracy
Şen, M. Umut, et al (2020) [5].	Real Life Deception Detection Dataset	Late fusion with visual, vocal and linguistic features	83%
Ahmed, Hammad Ud Din, et al (2021) [9]	Real Life Deception Detection Dataset	FACS with LSTM	89.4%
Mathur, Leena, and Maja J. Matarić (2020) [10]	Real Life Deception Detection Dataset	Visual and vocal features with Ada-Boost	84%
Krishnamurthy, Gangeshwar, et al (2018) [11]	Real Life Deception Detection Dataset	Early fusion with visual. vocal and linguistic features	96.1%
Yang, Jun-Teng, Guei-Ming Liu, and Scott C-H. Huang (2020) [20]	Real Life Deception Detection Dataset	EST, ME and IS13 features with logistic regression	92.7%
Proposed Method	**Real Life Deception Detection Dataset**	**Late fusion with visual, vocal and linguistic features**	**95.8%**

3.4 Drawbacks

Despite achieving great results, our approach has several significant drawbacks:

- The system relies on multiple deep learning models, which, when combined with the small dataset used, poses a risk of overfitting if not carefully managed.
- It is computationally intensive, taking approximately 2-3 minutes to process and predict a single sample.
- The models were specifically trained for courtroom environments, potentially reducing their accuracy in other settings.

4. CONCLUSION

The goal of this project was to investigate the boundaries of what could be accomplished in the field of deception detection using cutting edge deep learning and machine learning techniques. We drew inspiration for our work from recently published studies that achieved accuracy levels above 80%, including those in [5,], [9], [10], [11], and [20].

The primary tenet of the work was multimodal fusion; as a result, we began with singular modalities and achieved accuracy of 87.5% on video frames modality using 3D CNNs, 79.1% on audio data modality using bidirectional LSTMS, 79.1% on text data modality using ANN, and 75% on extracted micro-expressions modality using XGBoost. Following our work on individual modalities, we looked at multi-modal fusion, and our best model is late fusion (audio + video + text) using majority voting with 95.8% accuracy.

5. ACKNOWLEDGEMENTS

We would like to take this opportunity to express our gratitude to everyone who supported us throughout this project. We deeply appreciate their essential guidance, constructive feedback, friendly advice, and their genuine and insightful perspectives.

For her great assistance and support, we are especially appreciative of our supervisor, Dr. Salsabil Amin. Our profound gratitude also goes out to T.A. Zeina Rayan, whose guidance and support were invaluable throughout the process.

6. REFERENCES

[1] Bond Jr, C.F., DePaulo, B.M.: Accuracy of deception judgments. Personality and Social Psychology Review 10 (2006) 214–234.

[2] DePaulo, B.M., Lindsay, J.J., Malone, B.E., Muhlenbruck, L., Charl ton, K. and Cooper, H., 2003. Cues to deception. Psychological bulletin, 129(1),p.74.

[3] Perez-Rosas, V., Abouelenien, M., Mihalcea, R., Burzo, M.: Deception detection using real-life trial data. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM (2015).

[4] Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., Xiao, Y., Linton, C.J. and Burzo, M., 2015, September. Verbal and nonverbal clues for real-life deception detection. In Proceedings of the 2015 conference on empirical methods in natural language processing (pp. 2336-2346).

[5] ¸Sen, M.U., Perez-Rosas, V., Yanikoglu, B., Abouelenien, M., Burzo, M. and Mihalcea, R., 2020. Multimodal deception detection using real-life trial data. IEEE Transactions on Affective Computing, 13(1), pp.306-319.

[6] A. Vrij. Detecting Lies and Deceit: The Psychology of Lying and the Implications for Professional Practice. Wiley series in the psychology of crime, policing and law. Wiley, 2001.

[7] T . Gannon, A. Beech, and T. Ward. Risk Assessment and the Poly graph, pages 129–154. John Wiley and Sons Ltd, 2009.

[8] Bareeda, E.F., Mohan, B.S. and Muneer, K.A., 2021, May. Lie detection using speech processing techniques. In Journal of Physics: Conference Series (Vol. 1921, No. 1, p. 012028). IOP Publishing.

[9] Ahmed, H.U.D., Bajwa, U.I., Zhang, F. and Anwar, M.W., 2021. Deception detection in videos using the facial action coding system. arXiv preprint arXiv:2105.13659.

[10] Mathur, L. and Matarić, M.J., 2020, October. Introducing representations of facial affect in automated multimodal deception detection. In Proceedings of the 2020 International Conference on Multimodal Interaction (pp. 305-314).

[11] Krishnamurthy, G., Majumder, N., Poria, S. and Cambria, E., 2018, March. A deep learning approach for multimodal deception detection. In International Conference on Computational Linguistics and Intelli gent Text Processing (pp. 87-96). Cham: Springer Nature Switzer land.

[12] Rill-García, R., Jair Escalante, H., Villasenor-Pineda, L. and Reyes Meza, V., 2019. High-level features for multimodal deception detection in videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 0-0).

[13] S. Feng, R. Banerjee, and Y. Choi. Syntactic stylometry for deception detection. In Proceedings of the 50th Annual Meeting of the As sociation for Computational Linguistics: Short Papers - Volume 2, ACL '12, pages 171–175, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[14] Barathi, C.S., 2016. Lie detection based on facial micro expression body language and speech analysis. International Journal of Engineering Research & Technology, 5(2), pp.337- 343.

[15] A. Khaled, G. Walaa, M. Medhat, R. Alaa, and T. Ehab, "Lie Detection System," Ain Shams University, Faculty of Computer & Information Sciences, Scientific Computing Department.

[16] T. Yang, G.-M. Liu, and S. C.-H. . Huang, "Multimodal Deception Detection in Videos via Analyzing Emotional State-based Feature," arXiv:2104.08373 [cs], Apr. 2021.

[17] H. Karimi, J. Tang, and Y. Li, "Toward End-to-End Deception Detec tion in Videos," IEEE Xplore, Dec. 01, 2018.

[18] "Py-FEAT: Python Facial Expression Analysis Toolbox," GitHub, Feb. 23, 2022. license - cosanlab/py-feat · GitHub.

[19] V. Gupta, M. Agarwal, M. Arora, T. Chakraborty, R. Singh, and M. Vatsa, "Bag-of-Lies: A Multimodal Dataset for Deception Detection," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019.

[20] Goodfellow, I., Bengio, Y., and Courville, A., 2017. Deep Learning. MIT Press.

[21] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10), 1499-1503.

[22] Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001) (Vol. 1, pp. I-I). IEEE.

[23] Abdul, Z.K. and Al-Talabani, A.K., 2022. Mel frequency cepstral co efficient and its applications: A review. IEEE Access, 10, pp.122136 122158.

[24] Prince, E.B., Martin, K.B., Messinger, D.S. and Allen, M., 2015. Facial action coding system. Environmental Psychology & Nonverbal Behavior, 1.

[25] Pawłowski, M., Wróblewska, A. and Sysko-Romańczuk, S., 2023. Effective techniques for multimodal data fusion: A comparative analysis. Sensors, 23(5), p.2381.

[26] Bishop, C. M. (1995). Neural Networks for Pattern Recognition. Ox ford University Press.

[27] Cristianini, N., & Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press.

[28] Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.

[29] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[30] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Con ference on Knowledge Discovery and Data Mining (pp. 785-794).

[31] Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT'2010 (pp. 177-186). Physica-Verlag HD.

[32] Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing (3rd ed.). Prentice Hall.

[33] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.