Lie Detection via Facial and Vocal Analysis: An AI-Driven Application

Avram Andrei-Grigore
Faculty of Mathematics and Computer Science
Babeş-Bolyai University
Cluj-Napoca, Romania
andrei.grigore.avram@stud.ubbcluj.ro

***Abstract*—Lie detection remains a significant challenge in fields like security, law enforcement, and behavioral analysis. This paper presents an AI-based approach for detecting deception through the analysis of facial expressions and vocal characteristics. Initially, we explored multiple neural networks, incorporating 2D convolutional neural networks (CNNs) for static facial features, 3D CNNs for spatiotemporal visual analysis, and combined them with Emotion Recurrent Neural Networks (ERNNs) for vocal traits, trying a multimodal strategy. Although fusion techniques were investigated, empirical results revealed that the 3D CNN model alone provided the most reliable performance. Trained on both synthetic and real-world datasets, the final system achieved 84.05% accuracy. This paper details the experimental process, highlights the architectural evolution of the system, and evaluates the results, demonstrating the potential of deep learning in the automation of lie detection.**

***Index Terms*—Lie detection, deep learning, CNN, ERNN, facial expression analysis, voice analysis**

I. Introduction

Detecting deception has always been a complex and nuanced challenge across various fields, including criminal investigations, border control, national security, and corporate human resources. In high-stakes scenarios, the ability to determine whether someone is being truthful can play a pivotal role in decision-making. Traditional approaches such as polygraph examinations and expert behavioral analysis still hold value, but they require trained personnel, are time-consuming, and are often difficult to scale in fast-paced or remote environments.

Recent reports support the urgency of this problem. According to a 2023 report by the European Union Agency for Fundamental Rights (FRA), over 70% of border authorities expressed the need for improved tools to assess behavioral cues in real-time [1]. Additionally, a 2022 article in the journal *Nature Human Behaviour* indicates that humans are, on average, only 54% accurate in detecting lies—barely above random chance [2].

With the increasing availability of digital video and audio data, especially through online interactions, surveillance systems, and media platforms, there is an opportunity to develop automated systems capable of assisting in behavioral assessment. Advances in Artificial Intelligence (AI) and deep learning have made it feasible to train models that learn to recognize subtle behavioral patterns in both visual and auditory signals.

However, building such systems is not straightforward. Deceptive behavior rarely follows consistent or universal patterns. Unlike basic emotions, lying tends to vary significantly from person to person and is highly context-dependent. Another challenge lies in the limited availability of labeled datasets that contain authentic instances of deceptive behavior.

In this work, we explore the development of an AI-based system for lie detection by analyzing short video recordings of individuals while speaking. The original goal was to design a multimodal architecture that integrates facial expression and voice analysis using deep neural networks. We experimented with different combinations, including 2D Convolutional Neural Networks (CNNs) for static facial features, 3D CNNs for spatiotemporal modeling, and Enhanced Recurrent Neural Networks (ERNNs) for analyzing extracted vocal traits.

After extensive testing and evaluation, we found that a 3D CNN model trained solely on facial video sequences yielded the best and most consistent performance. This led to a refined implementation that focused exclusively on visual analysis using temporal convolution.

This paper documents the full development process, from initial experiments and model choices to dataset preparation and final performance evaluation, highlighting the strengths and challenges of using deep learning for automated deception detection.

II. Related Work

Before proposing a custom solution, it is essential to understand how deception detection has been addressed in recent literature. This chapter reviews various approaches based on visual, vocal, and multi-modal signals, highlighting key architectures and their respective performance.

A. Machine Learning Approaches

Numerous studies have explored the use of artificial intelligence and machine learning in the domain of deception detection. There exist several categories of approaches, each relying on different input modalities and feature extraction techniques. The most common branches are facial expression-based systems, vocal signal analysis, and multi-modal architectures combining multiple data sources.

Some studies focus on traditional machine learning algorithms applied to features extracted from facial and vocal data.

TABLE I: Comparison of Machine Learning Models

Models	Precision	Recall	F1-Score	Accuracy
KNN	0.85	0.85	0.85	0.85
SVM	0.93	0.93	0.92	0.92
Decision Tree (DT)	0.85	0.85	0.85	0.85
Random Forest (RF)	0.94	0.92	0.92	0.92
Naive Bayes (NB)	0.79	0.58	0.51	0.62
Extra Trees (ET)	0.94	0.92	0.92	0.92

TABLE II: Per Participant Model Performance

Metric	p1	p2	p3	p4	p5	p6	p7	p8	p9
ACC	46.77	75.00	45.75	79.88	87.97	56.25	62.25	74.38	24.63
F1	29.79	53.12	30.48	45.65	50.00	9.01	38.72	51.90	26.44
Metric	p10	p11	p12	p13	p14	p15	p16	p17	**ALL**
ACC	45.38	**97.88**	93.25	62.88	46.25	51.38	92.00	46.67	65.00
F1	13.01	**65.75**	63.33	51.90	4.38	4.44	59.47	20.89	63.12

In the paper [3], the authors propose a deception detection system based on facial and vocal features using Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbor (KNN). Random Forest achieved the best performance, surpassing 90% accuracy. A comparison between these machine learning models is shown in Table I.

A comprehensive review [4] analyzed over 80 papers and reported average SVM accuracy around 84%. Another paper [5] introduced a personalized card game dataset and tested machine learning models both generally and individually. While general models had modest accuracy (57.4%), specialized models trained for each individual reached up to 97.8% accuracy, as shown in Table II.

B. Deep Learning Techniques

Deep learning techniques have also been extensively applied. In the article [6], an Enhanced Recurrent Neural Network (ERNN) model combining Long Short-Term Memory (LSTM) and fuzzy logic achieved 97.3% accuracy and 97.77% F1-score, outperforming Artificial Neural Network (ANN) and classical Recurrent Neural Networks (RNNs).

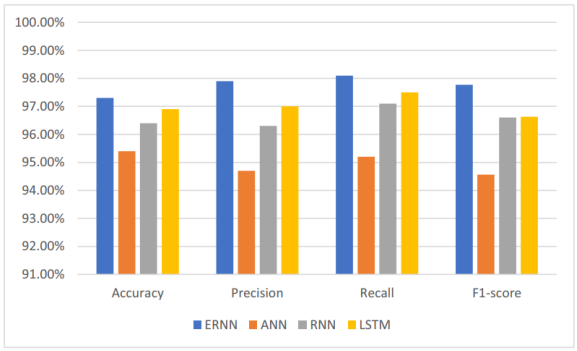

Fig. 1: Comparison between ERNN and other models (ANN, RNN, LSTM). Source: [6].

Another study [7] combined Convolutional Neural Networks (CNNs) with LSTMs to model temporal behavior using facial features such as yaw, pitch, and microexpressions.

Similarly, the paper [8], focused on microexpression detection using 3D CNNs trained on a large video dataset, achieving over 93% accuracy. These results highlight that purely visual models can be effective under controlled conditions.

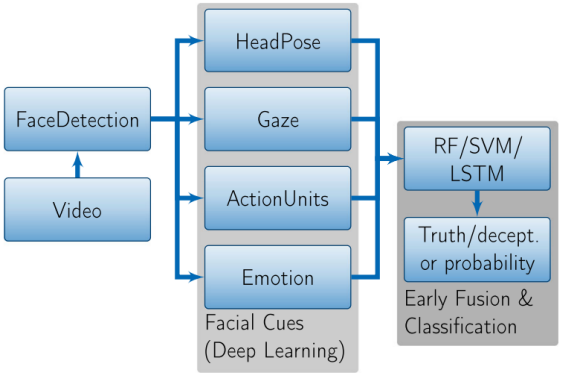

Fig. 2: Proposed CNN-LSTM architecture for temporal facial behavior analysis. Source: [7].

Facial expression-only systems were also examined. The work in [9] tested OpenFace and Dense Trajectories, where OpenFace combined with SVM achieved a 0.78 AUC, outperforming human baseline performance. Additional methods used pre-trained CNNs and AdaBoost to extract and classify facial affective features [10]. A review [11] discussed limitations in generalizability due to sparse datasets and inconsistent facial signals.

C. Audio-Based Approaches

In the audio domain, the study in [12] used Mel-Frequency Cepstral Coefficients (MFCC) features and LSTM for vocal stress detection, reaching 97.3% accuracy. The overall architecture used in this study is presented in Figure 3.

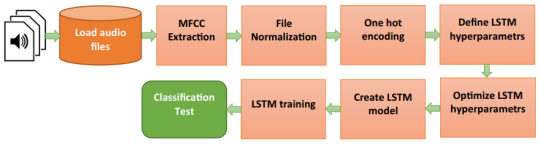

Fig. 3: ERNN architecture for vocal stress analysis. Source: [12].

D. Multimodal Approaches

Multimodal systems that combine visual and audio channels show the most promise. The approach in [13] integrated 3D CNN with OpenSmile features, comparing fusion at the feature and decision levels, achieving up to 100% true-positive rate. Another review [14] highlighted multiple fusion strategies, with some achieving as high as 97.6% accuracy.

Additional work such as [15] and [16] explored Transformer-based multimodal learning and modular fusion pipelines, respectively. These architectures are capable of

leveraging partial or incomplete data (e.g., missing audio or video) while maintaining high accuracy.

Finally, a large-scale dataset introduced in [17] enabled the training of robust multimodal models that achieved over 95% accuracy, confirming the scalability of visual-audio deception detection.

These studies illustrate the wide variety of approaches to deception detection. While multimodal systems offer the best performance, simpler models such as 3D CNNs can still be competitive when applied to well-processed visual data alone. At the same time, limitations in dataset diversity, generalizability, and explainability remain open research challenges. These insights form the foundation for our own experimental attempts, discussed in the next section.

III. Experimental Attempts

Building upon insights from previous studies, this chapter presents a series of practical experiments using different model architectures and datasets. The goal was to evaluate which components contribute most to deception detection performance.

A. Facial-Based 3D CNN Model

In the early stages of model development, a 3D Convolutional Neural Network (3D CNN) architecture was implemented for facial analysis. This model processed both spatial and temporal information, allowing it to capture micro-expressions and subtle behavioral changes over time. The input consisted of short video sequences, each transformed into consecutive frames, from which facial regions were cropped and normalized. The dataset initially used was the Real-life Trial Deception Detection Dataset [18], containing courtroom footage labeled as truthful or deceptive.

The training and validation curves for this 3D CNN model, shown in Figure 4, demonstrate a stable learning process. However, the confusion matrix in Figure 5 highlights misclassifications, indicating difficulty in distinguishing between truthful and deceptive behavior.

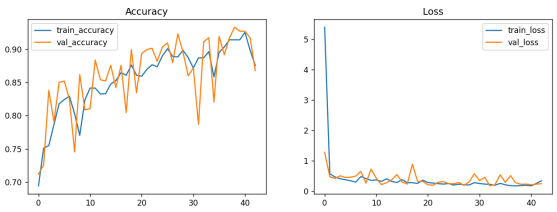

Fig. 4: Training and validation curves for the 3D CNN on the trial dataset.

B. Frame-Based 2D CNN Model

Despite these seemingly satisfactory results, the model's performance on new, self-recorded videos was inconsistent and generally poor, revealing a lack of generalization. This limitation led to the exploration of a simpler 2D CNN, designed to analyze individual frames independently. The training and validation performance of this model is illustrated in Figure 6, and its prediction distribution on the validation set is detailed in the confusion matrix shown in Figure 7.

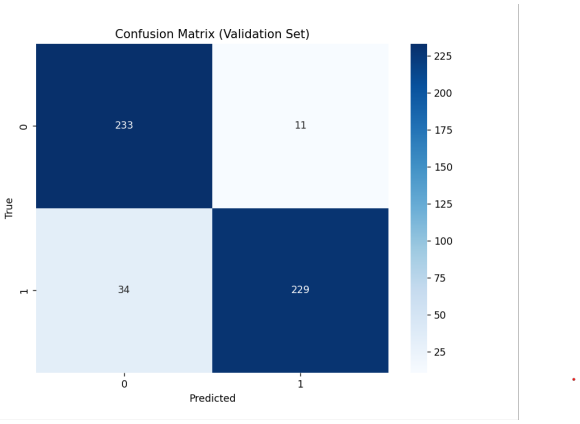

Fig. 5: Confusion matrix for 3D CNN on the trial dataset.

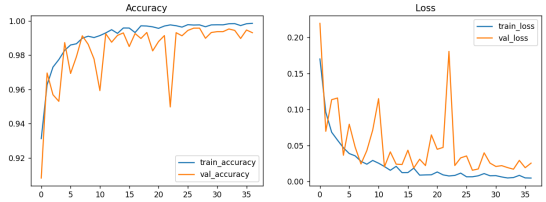

Fig. 6: Training and validation curves for the 2D CNN on the trial dataset.

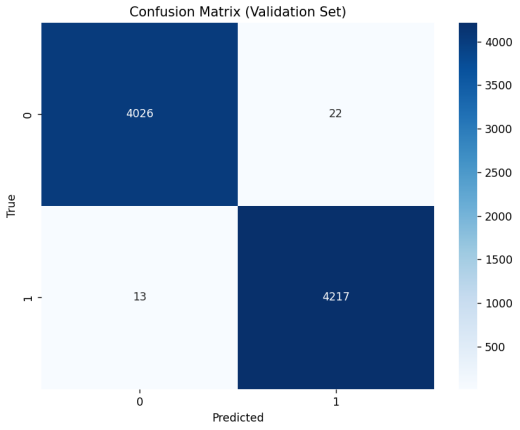

Fig. 7: Confusion matrix for 2D CNN on the trial dataset.

C. Audio-Based ERNN Model

To further enhance model performance, an Enhanced Recurrent Neural Network (ERNN) was trained on audio features extracted from the same courtroom dataset. The training process, depicted in Figure 8, showed good convergence. Nevertheless, as seen in the corresponding confusion matrix (Figure 9), the

model still struggled with classification accuracy on unseen, real-world data.

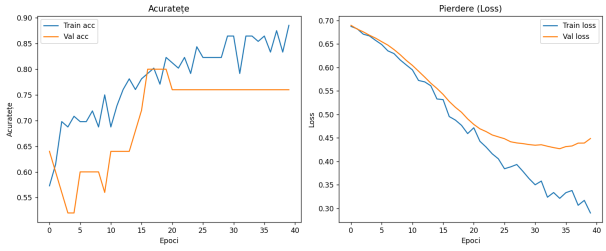

Fig. 8: Training curves for ERNN vocal model.

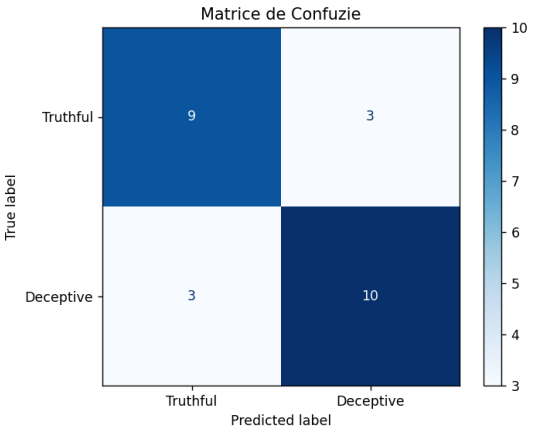

Fig. 9: Confusion matrix for ERNN vocal model.

D. Dataset Expansion and Generalization Limitations

To diversify the data and improve results, a second dataset containing facial micro-expressions in day-to-day contexts was incorporated [19]. This dataset added variation in expressions and environments, offering broader learning opportunities. However, even with the inclusion of this additional data, none of the implemented models (3D CNN, 2D CNN, or ERNN) were able to consistently deliver accurate predictions on real-world inputs.

E. Hardware and Software Environment

All experiments were conducted on a local workstation equipped with an 11th Generation Intel(R) Core(TM) i5-11400H CPU @ 2.70GHz, with 16 GB RAM (15.7 GB usable), running a 64-bit operating system on an x64-based processor architecture. No GPU acceleration was used in the final phase of experimentation. For these experiments, we used Python language, and libraries such as Tensorflow, Keras, Mediapipe, Numpy, Pandas, and some others.

In conclusion, the experiments conducted so far provided useful insights into model behavior and generalization challenges. These findings highlighted the limitations of the current approaches and served as motivation to refine the overall architecture and processing pipeline. The next section introduces the final system proposed to address these issues more effectively.

IV. Final Solution

After analyzing the limitations of the previously tested models, a refined and focused approach was adopted. The final system is built exclusively on a 3D CNN architecture, optimized to improve generalization and robustness when analyzing human facial expressions from video sequences.

The complete pipeline begins with video input, from which frames are extracted and facial regions are detected using the MediaPipe face detection algorithm. These facial crops are resized to 64×64 pixels, normalized, and then grouped into sequences of 16 consecutive frames to form valid 3D inputs.

The 3D CNN architecture consists of three blocks, each including a `Conv3D` layer followed by `BatchNormalization` and `MaxPooling3D`. The first convolutional kernel was adjusted to `(1,2,2)` to better preserve temporal information across adjacent frames. Following the convolutional stages, a `Flatten` layer feeds into a `Dense(256)` with ReLU activation, and the output is generated through a `Dense(2)` with softmax activation for binary classification. The complete structure of the final 3D CNN model used in our solution is visualized in Figure 10, summarizing all layers and their dimensions.

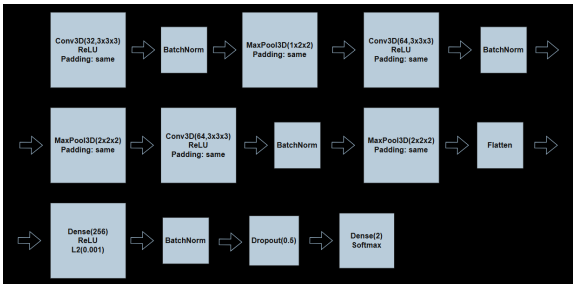

Fig. 10: Final 3D CNN model architecture.

The training was performed using categorical cross-entropy loss and the Adam optimizer. To enhance the model's generalization capabilities, data augmentation was applied frame-wise, and class imbalance was handled using class weighting. Additionally, callbacks such as `EarlyStopping`, `ReduceLROnPlateau`, and `ModelCheckpoint` were incorporated.

The model achieved stable training behavior and strong classification performance, as evidenced by both the learning curves and evaluation metrics.

As shown in Figure 11, the training and validation accuracy steadily increased throughout the 50 training epochs, eventually stabilizing at values above 85%. The loss curves similarly decreased, indicating a consistent optimization process with minimal signs of overfitting. Notably, the validation accuracy closely tracks the training accuracy, which suggests good generalization to unseen data. This implies that the model learned

relevant temporal-spatial patterns from the video sequences without memorizing training samples.

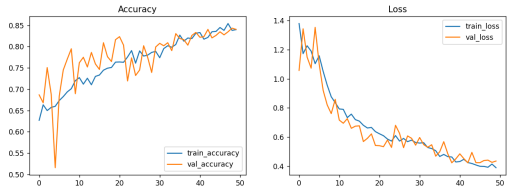

Fig. 11: Training and validation curves for the final 3D CNN model.

Further analysis using the confusion matrix (Figure 12) provides deeper insights into the model's prediction capabilities. Out of the total validation samples, 335 truthful cases (label 0) were correctly classified, while only 27 were misclassified as deceptive. In contrast, 255 deceptive cases (label 1) were correctly detected, with 85 being misclassified as truthful. These results indicate that the model has a relatively higher sensitivity (true positive rate) for detecting truthful behavior, but still maintains competitive performance in identifying deception.

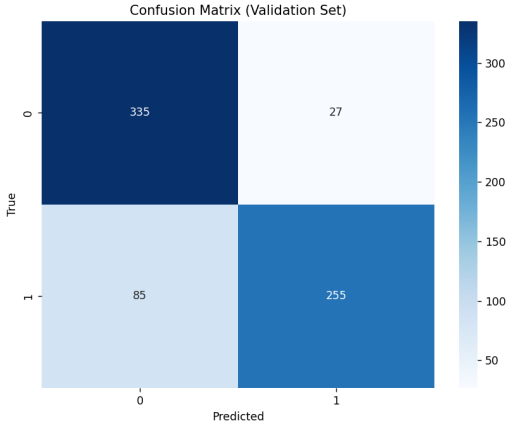

Fig. 12: Confusion matrix for the final 3D CNN model.

The balanced classification outcome and low false-positive rate support the suitability of 3D CNNs for learning both microexpressions and contextual facial cues across time. These findings reinforce the idea that temporal convolutional models can serve as a strong foundation for video-based deception detection, especially when trained on diverse and well-preprocessed datasets.

The performance summary of the model is shown in Figure 13, where an accuracy of 84.05%, precision of 84.93%, recall of 84.05%, and F1 score of 83.89% were recorded.

To provide a comprehensive understanding of the system's functionality, the processing pipeline is illustrated in Figure 14. It describes the steps from video input to prediction output.

Overall, the final solution demonstrates a significant improvement in robustness and interpretability compared to earlier attempts. It sets a solid foundation for future research in multimodal deception detection.

Metric	Value
Accuracy	0.8405
Precision	0.8493
Recall	0.8405
F1 Score	0.8389

Fig. 13: Final performance metrics of the 3D CNN model.

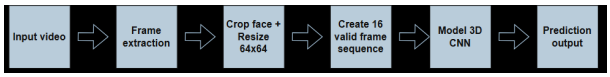

Fig. 14: Complete processing pipeline for the final 3D CNN model.

Even if the presented results demonstrate clear improvements and validate the decision to streamline the system, challenges remain, which we discuss in the following section covering limitations and future directions.

In the next chapter, we place our results in context by comparing them with prior work in the literature, highlighting similarities, distinctions, and performance differences in various conditions.

V. Comparisons and Discussion

After implementing our final 3D CNN-based solution, we compared its results to those reported in related works. While our model achieved 84.05% accuracy, several studies reported higher scores under controlled conditions. For instance, [6] and [12] demonstrated over 97% accuracy using audio-based ERNN models. Similarly, the multimodal fusion method described in [13] reached a perfect true-positive rate by combining OpenSmile and 3D CNN features.

However, a key distinction lies in the source and nature of data. Many existing works rely on curated datasets with constrained conditions, limited demographic diversity, or even synthetically induced deception. In contrast, our approach focused on generalization from courtroom videos and real-life self-recorded sequences. This trade-off reflects the broader research challenge: models with high accuracy often underperform when exposed to real-world unpredictability.

Moreover, while our model avoids voice data in its final form, future improvements could include re-integrating multimodal fusion. Our work can be positioned between the robust but narrow methods seen in [9] and more ambitious frameworks such as in [15] which use Transformers for flexible-modal deception detection.

The results also indicate that using only facial features can be competitive. Compared to [10], who reported 79% accuracy with multimodal AdaBoost classifiers, our purely visual model performs comparatively well. This validates our pipeline's effectiveness and supports its applicability in settings where audio capture is limited or undesirable.

Even if the presented results demonstrate clear improvements and validate the decision to streamline the system, challenges remain, which we discuss in the following section covering limitations and future directions.

VI. Limitations and Future Work

Despite the improved performance of the final 3D CNN-based system, several limitations remain that affect its broader applicability in real-world settings. First, the generalization of the model outside the training domain is still a challenge. Although the system performs well on validation data, its behavior on entirely new, uncontrolled environments is unpredictable. This is largely due to the restricted diversity of the datasets used, most video recordings coming from limited scenarios such as courtroom footage or controlled interviews.

Another important limitation is the absence of vocal and textual information in the final model. Earlier attempts at including vocal cues (via ERNN) were not effective due to variability in speaker characteristics and insufficient data volume. However, completely excluding this modality leaves out potentially valuable information.

For future work, several directions are worth exploring. One key improvement would be the extension of the dataset to include more diverse video samples from various real-world situations such as interviews, online vlogs, or casual conversations in natural conditions. This could help the model better learn deception patterns that generalize across contexts.

Additionally, reintroducing the vocal modality, perhaps in combination with speech-to-text transcription for semantic analysis, may provide a more comprehensive understanding of deceptive behavior. Combining facial expressions, vocal tone, and spoken content could allow the model to analyze contradictions between what is said and how it is said.

Another promising direction involves experimenting with transformer-based architectures that can simultaneously process multi-modal data such as images, audio, and text. These models have demonstrated strong capabilities in other tasks and could outperform current CNN-based architectures in this domain.

Finally, integrating explainable AI (XAI) methods into the system would greatly improve user trust and transparency. For example, visual heatmaps showing which parts of the face influenced the model's prediction could make the decision-making process more interpretable and justifiable in practical applications.

While the current model represents a solid step forward, it opens up a wide range of possibilities for extension, both in terms of data richness and modeling complexity.

Addressing these limitations could greatly enhance the system's applicability. The final section of this paper will summarize the entire development process and key contributions.

VII. Conclusion

After exploring various experimental paths and refining the approach, this final section consolidates the insights and outcomes of the study.

This paper presented the design and development of an AI-based system for detecting deception through facial analysis. Starting with multi-modal experiments—including 3D CNNs, 2D CNNs, and ERNN, we observed that initial architectures performed well on validation sets but failed to generalize to real-world, self-recorded scenarios.

To address this, a second dataset was introduced and numerous adjustments were made to the preprocessing pipeline, especially in terms of face detection and temporal coherence. These experiments highlighted key challenges in cross-context lie detection and offered critical insights into model behavior.

The final system was built around a focused and optimized 3D CNN architecture, which integrated improved frame sampling, MediaPipe face detection, sequence formation, and class balancing. This model achieved an accuracy of 84.05% and proved more robust in diverse test conditions.

While the preceding section outlined current limitations and promising future directions—including multimodal integration and explainability—this work lays a strong foundation for continued research in deception analysis using visual cues. It demonstrates that, with the right architectural and preprocessing choices, deep learning can approach the subtle and complex task of lie detection in an interpretable and scalable way.

References

[1] European Union Agency for Fundamental Rights, "Fundamental rights at borders: Annual report 2023," 2023, accessed: 2025-06-24. [Online]. Available: https://fra.europa.eu/en/publication/2023/fundamental-rights-border-2023

[2] C. F. Bond and B. M. DePaulo, "The truth about lying: A meta-analytic review of deception detection accuracy," *Nature Human Behaviour*, vol. 6, no. 8, pp. 1045–1053, 2022.

[3] X. Wang, Y. Wu, and Y. Zhang, "Multimodal deception detection based on facial micro-expressions and acoustic features," *Journal of Intelligent & Fuzzy Systems*, vol. 45, no. 3, pp. 3575–3586, 2023.

[4] R. de Oliveira Anacleto, M. Kalil, and J. P. M. de Oliveira, "Deception detection with machine learning: A systematic review," *PLOS ONE*, vol. 18, no. 2, p. e0281323, 2023.

[5] L. Rodriguez-Diaz, J. Carbo, K. A. F. Mora, and J.-M. Odobez, "Deception detection through personalized deep learning models on face and voice," in *2022 17th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2022, pp. 1–8.

[6] M. El-Attar, "Explainable enhanced recurrent neural network for lie detection using voice stress analysis," *Multimedia Tools and Applications*, vol. 83, no. 1, pp. 32 277–32 299, 2024.

[7] L. Dinges, M.-A. Fiedler, A. Al-Hamadi, T. Hempel, A. Abdelrahman, J. Weimann, D. Bershadskyy, and J. Steiner, "Exploring facial cues: automated deception detection using artificial intelligence," *Neural Computing and Applications*, vol. 36, pp. 14 857–14 883, 2024.

[8] M. Osama and Z. Zahid, "Hybrid lie detector: Using facial micro-expressions and physiological data," 2024, undergraduate thesis, NUST Balochistan Campus, Pakistan.

[9] 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, p. 107063, 2022.

[10] D. Mathur and V. Shah, "Introducing representations of facial affect in automated multimodal deception detection," *arXiv preprint arXiv:2008.13369*, 2020.

[11] 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.

[12] F. M. Talaat, "Explainable enhanced recurrent neural network for lie detection using voice stress analysis," *Multimedia Tools and Applications*, vol. 83, pp. 32 277–32 299, 2024.

[13] S. Chebbi and S. B. Jebara, "Deception detection using multimodal fusion approaches," *Multimedia Tools and Applications*, vol. 82, pp. 13 073–13 102, 2023.

[14] S. L. King and T. Neal, "Applications of ai-enabled deception detection using video, audio, and physiological data: A systematic review," *IEEE Access*, vol. 12, pp. 135 207–135 229, 2024.

[15] S. Li, Y. Zhao, and J. Zhang, "Flexible-modal deception detection with audio-visual adapter," *arXiv preprint arXiv:2302.05727*, 2023.

[16] A. Srivastava, "Enhancing lie detection in video-based speech through machine learning," *The National High School Journal of Science*, vol. 2025, no. Spring, 2025, available online via NHSJS.

[17] Y. Guo, M. Sun, Z. Li, Y. Li, S. Zhao, L. Yang, and Z. Li, "Audio-visual deception detection: Dolos dataset and parameter-efficient crossmodal learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, link accesed at April 10, 2025. [Online]. Available: https://arxiv.org/abs/2303.12745

[18] V. Perez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 59–66.

[19] D. Mathur, "Micro expression dataset for lie detection," 2024, accessed on May 10, 2025. Available at https://www.kaggle.com/datasets/devvratmathur/micro-expression-dataset-for-lie-detection/data.