IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AI-BASED MULTIMODAL SYSTEM FOR TRUTH AND LIE DETECTION

Deeksana K, Sivakrishnan M, Varnamalika N, Mrs. P. Prema

Student, Student, Student, Faculty
Department of Artificial Intelligence and Data Science,
Sri Venkateswara College of Engineering, Chennai, India

DOI: https://doi.org/10.56975/ijcrt.v14i7.309010

***Abstract*:** This paper presents an AI-based multimodal system for truth and lie detection that combines audio and video analysis within a unified Flask based framework for non-intrusive deception assessment. Traditional lie detection techniques, such as polygraph tests, are often intrusive, expensive, and not always reliable in real-world scenarios. To overcome these limitations, the proposed system makes use of recent advancements in computer vision and speech processing to analyze both facial behavior and vocal characteristics. The audio module extracts deep speech features along with handcrafted features such as pitch, energy, and zero-crossing rate, while the video module uses a CNN-LSTM architecture to capture spatial and temporal patterns from facial expressions and movements. The outputs from both modules are combined using a multimodal fusion approach to produce a final prediction along with a confidence score. The system supports both file-based inputs and real-time recording through a user-friendly web interface, and it also maintains a history of predictions for better usability. Although previous studies show that multimodal approaches generally perform better than single-modality systems, challenges such as limited datasets, generalization issues, and ethical concerns still remain. Overall, the proposed system provides a practical and scalable approach for developing AI-assisted deception detection applications.

Keywords: Audio classification, CNN-LSTM, Wav2Vec2 Model, Deception detection, Multimodal learning.

I. INTRODUCTION

The rapid growth of artificial intelligence and machine learning has significantly transformed the way human behavior is analyzed and interpreted. One important application of these technologies is deception detection, which plays a critical role in areas such as security, forensic investigation, and behavioral analysis. Traditionally, lie detection has relied on methods such as polygraph tests, which measure physiological signals like heart rate and skin conductivity. However, these methods are often intrusive, require specialized equipment, and may not always provide reliable results.

In recent years, research has shifted toward automated deception detection systems that utilize visual and audio data. Facial expressions, micro-expressions, eye movements, and speech patterns are considered important indicators of deceptive behavior. However, most existing systems rely on a single modality, such as either video or audio, which limits their effectiveness. Human behavior is inherently complex, and relying on only one type of input often leads to reduced accuracy and poor generalization across different individuals and environments.

To address these limitations, multimodal approaches have gained attention, as they combine information from multiple sources to improve prediction performance. By integrating both visual and acoustic features, these systems can capture a more comprehensive representation of human behavior. However, challenges such as subtle micro-expressions, variability in facial cues among individuals, environmental noise, and real-time processing constraints still remain significant obstacles.

This project proposes an AI-based multimodal system for truth and lie detection that leverages both video and audio inputs for improved accuracy and robustness. The system employs deep learning techniques such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for extracting spatial and temporal features from video data, along with audio feature extraction methods like Mel Frequency Cepstral Coefficients (MFCC) and advanced models such as Wav2Vec2 for speech analysis. The extracted features are fused and processed through a classification model to predict whether the input corresponds to truth or deception, along with a confidence score.

The main objective of this work is to develop a non-intrusive, accurate, and scalable deception detection system that overcomes the limitations of traditional and unimodal approaches. By combining multimodal data and advanced machine learning techniques, the proposed system aims to provide a more reliable and practical solution for real-world applications.

II. LITERATURE REVIEW

Delmas et al. (2024), "A Review of Automatic Lie Detection from Facial Features"

[1]. This study provides a comprehensive review of 28 research works focused on deception detection using facial features. The authors analyze various datasets, feature extraction techniques, and classification models used in prior studies, while also evaluating their theoretical foundations based on the leakage hypothesis. The findings indicate that the average accuracy of such systems ranges between 61.87% and 72.93%, highlighting the limitations of relying solely on facial cues. The relevance of this study lies in its critical observation that most existing approaches lack strong theoretical grounding and fail to generalize effectively. This insight emphasizes the need for integrating multiple modalities rather than depending only on facial expressions, which directly supports the motivation behind the proposed multimodal system.

Tsuchiya, Hatano, and Nishiyama (2023), "Detecting Deception Using Machine Learning with Facial Expressions and Pulse Rate"

[2]. This work introduces a multimodal approach combining facial expression analysis with physiological signals such as pulse rate. The dataset was collected in realistic conditions using webcams and wearable devices, and a Random Forest classifier was employed for prediction. The system achieved accuracy and F1-scores in the range of 0.75–0.88, demonstrating improved performance compared to unimodal systems. The importance of this study lies in its demonstration that combining multiple modalities enhances deception detection performance. It also highlights that deceptive behavior varies across individuals, suggesting the need for adaptable and robust models, which influenced the design of the proposed system.

Dinges et al. (2024), "Exploring Facial Cues: Automated Deception Detection Using Artificial Intelligence"

[3]. This research focuses on extracting facial features such as expressions, gaze, and head pose using convolutional neural networks. The study applies early fusion techniques to combine multiple facial cues and evaluates the model across different datasets. The results indicate that facial expressions are more effective than gaze or head pose individually, while multimodal feature fusion consistently improves accuracy. The contribution of this work is its validation of feature fusion strategies in improving detection performance. This supports the idea of integrating diverse behavioral cues, which is extended further in the proposed system by incorporating both visual and audio modalities.

Rote, Shivama, and Narayanamoorthi (2025), "Lie Detection Using Audio Classification"

[4]. This study explores deception detection using audio data, specifically leveraging convolutional neural networks to automatically extract features from speech signals. The model was trained on real-world courtroom audio data and achieved an accuracy of 93.5%, demonstrating the effectiveness of audio-based analysis. The significance of this paper lies in its emphasis on automated feature extraction from audio, eliminating the need for manual engineering. It highlights the potential of speech-based cues such as tone, pitch, and stress, which are incorporated into the proposed system using MFCC and advanced audio processing techniques.

Chaskar et al. (2025), "Lie Detection Using Facial Expressions and Speech Analysis"

[5]. This work proposes a multimodal framework combining facial expression analysis with speech stress analysis using the RAVDESS dataset. The system processes audio and video data separately and integrates the results using a unified neural network. The findings reveal that combining both modalities significantly improves accuracy compared to single-modal systems. The relevance of this study is its strong support for multimodal integration. It demonstrates that emotional and behavioral cues from both speech and facial expressions are essential for reliable deception detection, directly aligning with the core concept of the proposed system.

Galli et al. (2026), "Truth or Lie: An Audio-Visual Approach to Deception Detection"

[6]. This research presents an advanced audio-visual framework that extracts temporal patterns from both acoustic features and facial Action Units. The system employs a late fusion strategy combined with a meta-learning module to enhance prediction accuracy. The results show that the audio-visual approach outperforms both unimodal and earlier multimodal methods. The key contribution of this work is its demonstration of the importance of temporal dynamics and multimodal fusion in real-world scenarios. It also highlights the feasibility of non-intrusive, AI-based deception detection systems, which directly supports the design and objectives of the proposed project

III.METHODOLOGY

The proposed system follows a structured and modular pipeline designed to accurately detect deception using both visual and audio information. Each stage in the methodology plays a critical role in transforming raw input data into meaningful predictions. The complete workflow, as illustrated in Figure 1, ensures efficient data handling, robust feature extraction, and reliable classification by integrating multimodal learning techniques.

3.1 Data Acquisition and Preprocessing

The first stage involves collecting input data in the form of video and audio, either through real-time recording or from prerecorded datasets. Since raw data is often noisy and unstructured, preprocessing is essential to ensure consistency and improve model performance.

For the **video modality**, the input is divided into frames at a fixed frame rate. Each frame undergoes face detection to identify the region of interest, followed by face alignment to correct pose variations. The faces are then resized and normalized to maintain uniformity. Data augmentation techniques such as flipping, rotation, and brightness adjustment are also applied to increase diversity and reduce overfitting.

For the **audio modality**, the speech signal is extracted and converted into a standard format (16 kHz, mono). It is then segmented into smaller time windows to capture detailed variations. Additional steps like noise reduction and normalization help improve audio quality. Overall, preprocessing ensures that both audio and visual data are clean and suitable for accurate analysis.

3.2 Feature Extraction

This stage focuses on extracting meaningful features from both video and audio using deep learning techniques. For the **visual stream**, a CNN-LSTM model is used. The CNN extracts spatial features such as facial expressions, eye movements, and muscle changes from individual frames. These features are then passed to an LSTM network, which captures temporal patterns across frames, helping to detect micro-expressions and subtle changes over time.

For the **audio stream**, feature extraction targets speech characteristics like pitch, tone, and frequency. Techniques such as MFCC are used to represent speech signals, while advanced models like Wav2Vec2 learn deeper patterns directly from raw audio. These help identify cues like hesitation, stress, and voice modulation.

By combining both handcrafted features (MFCC) and deep learning representations (Wav2Vec2), the system effectively captures important audio and visual cues for deception detection.

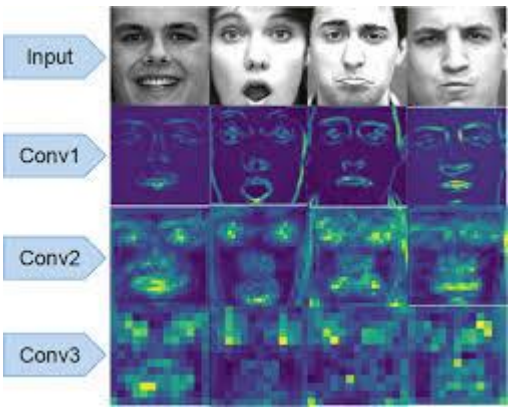

Figure 1: Feature Extraction

3.3 Feature Fusion and Classification

After extracting features from both audio and video modalities, the next step is to integrate them into a unified representation. This process, known as multimodal fusion, plays a key role in combining complementary information from speech and visual cues to improve prediction reliability.

In this system, the predictions obtained from both pipelines are combined using a decision-level (late) fusion approach. The outputs from the audio and video models are integrated to generate a final prediction, allowing the system to utilize both speech-based features such as tone and pitch, and visual features such as facial expressions and temporal changes. This approach ensures flexibility and robustness in handling different input conditions.

The fused output is then used to determine the final classification as truth, lie, or uncertain, along with a confidence score. By combining both modalities, the system reduces dependency on any single input source and improves overall accuracy and generalization. This multimodal strategy enables more reliable detection, especially in real-world scenarios where one modality may be weak, noisy, or unavailable.

3.4 Result Interpretation

In this stage, the system generates the final prediction based on the outputs obtained from the audio and video analysis along with the multimodal fusion process. The combined information is used to determine whether the input corresponds to truthful or deceptive behavior.

The final result is displayed as “Truth” or “Lie,” along with a confidence score that indicates the certainty of the prediction. This confidence score helps users understand how reliable the result is. Additionally, in cases where the input is unclear or noisy, the system may output an “uncertain” result. This approach improves transparency, reliability, and usability of the system in real-world applications.

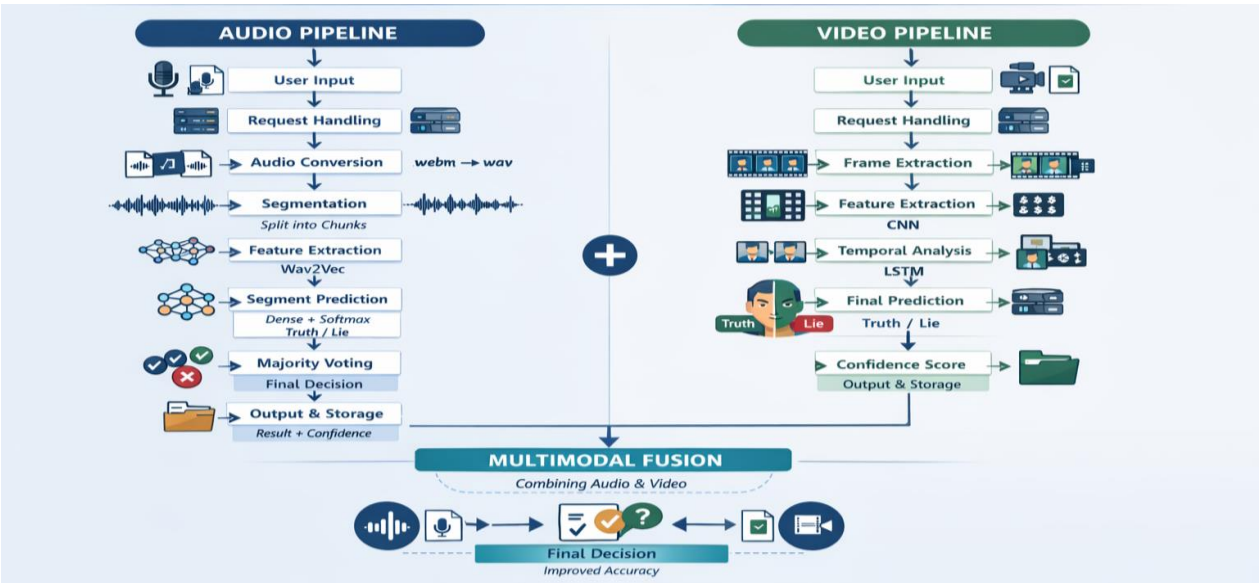

Figure 2: Work-architecture diagram

IV. RESULTS

The proposed system processes both video and audio inputs, which are divided into training, validation, and testing datasets for proper evaluation.

During training, the model learns to identify subtle behavioral patterns such as facial micro-expressions from video data and speech variations from audio signals. The video data is analyzed using CNN-LSTM models, while audio data is processed using MFCC and deep learning-based approaches. The model captures both spatial and temporal features that are important for detecting deception.

The system evaluates performance using common metrics such as accuracy, precision, recall, and F1-score. The results show that the multimodal approach improves detection performance compared to using a single modality. The classification results are further analyzed using a confusion matrix, which demonstrates the model's ability to correctly identify both truthful and deceptive inputs.

In addition to overall prediction, the system performs segment-level analysis, where predictions are generated for smaller portions of the input. A final decision is obtained using majority voting, and the output is presented along with a confidence score.

Interpretability is an important aspect of this system. Instead of providing only a binary result, the model outputs a confidence percentage, which helps in understanding the reliability of the prediction. This improves transparency and makes the system more suitable for real-world applications.

Figure 3: Multimodal Lie Detection UI

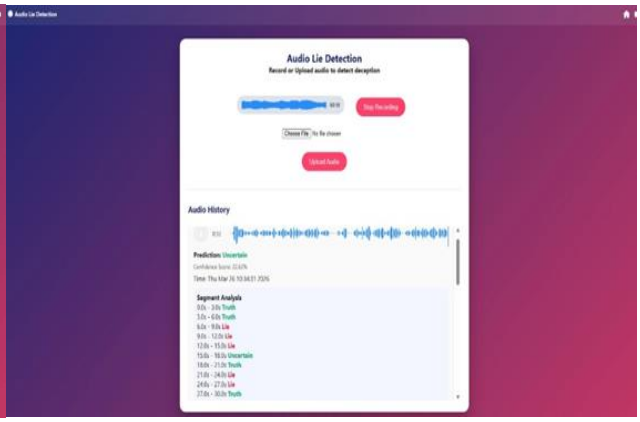

Figure 4: Live audio recording for detection

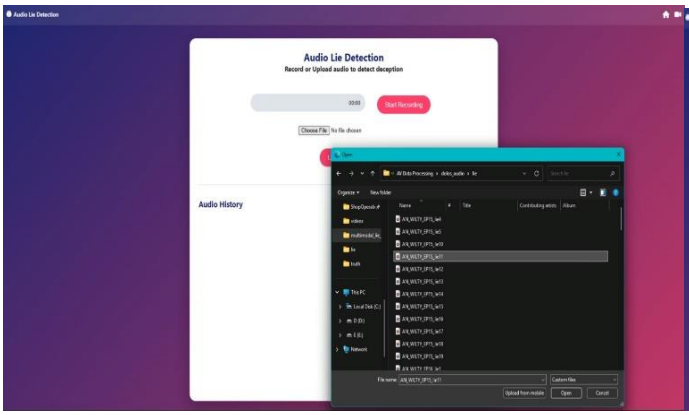

Figure 5: Uploading audio file for detection

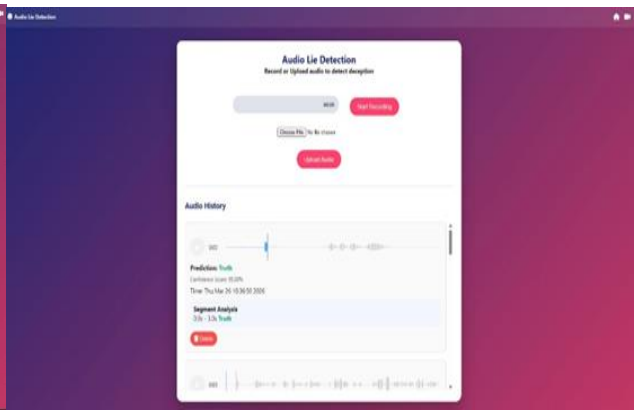

Figure 6: Live & Upload Results

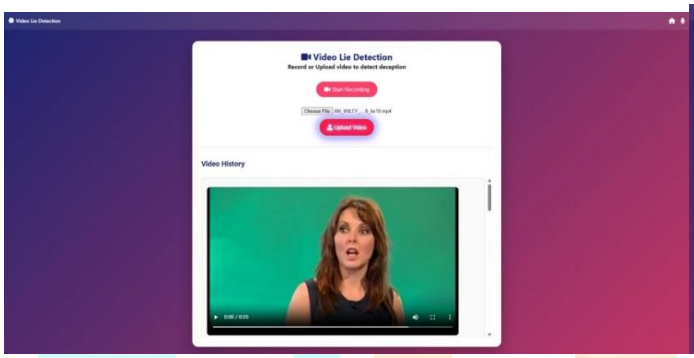

Figure 7: Video Lie Detection

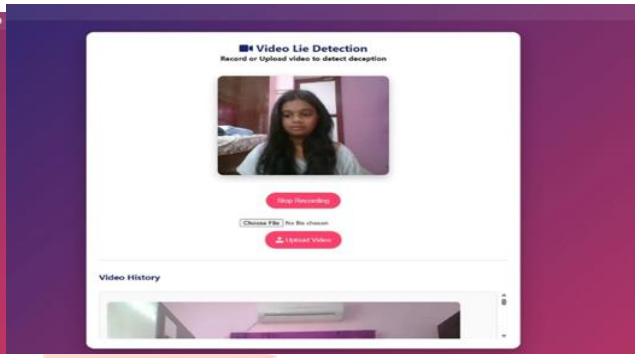

Figure 8: Live video recording for detection

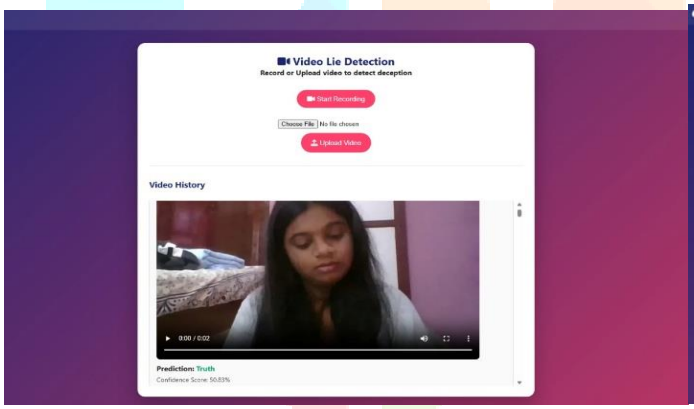

Figure 9: Live Results

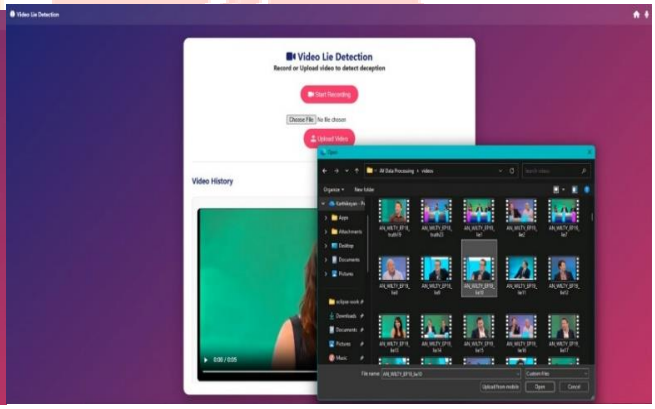

Figure 10: Uploading video file for detection

.

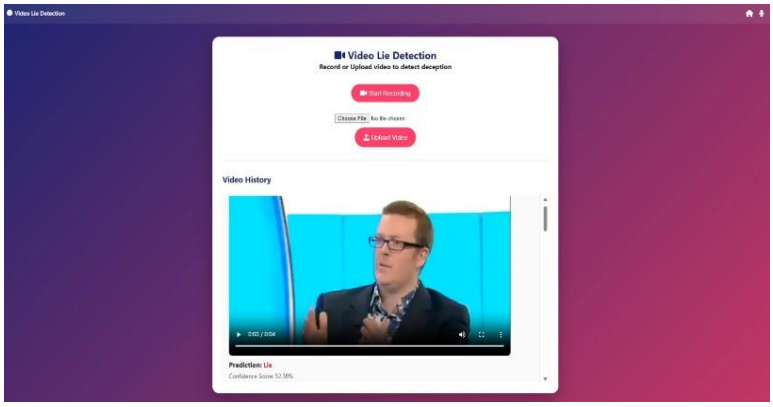

Figure 11: Live & Upload Results

V. APPLICATION

The proposed multimodal truth and lie detection system has various applications in different areas:

- **Security and Investigation** – It can assist law enforcement agencies in identifying deceptive behavior during interrogations and investigations.
- **Forensic Analysis** – It can help in analyzing statements and testimonies to determine their authenticity in legal cases.
- **Recruitment and Interviews** – It can be used in hiring processes to evaluate candidate responses and detect potential deception.
- **Online Proctoring** – It can monitor students during online examinations to detect suspicious or dishonest behavior.
- **Customer Service and Feedback Analysis** – It can analyze customer interactions to identify genuine and misleading feedback.
- **Human-Computer Interaction** – It can enhance intelligent systems by enabling them to understand user behavior and respond accordingly.
- **Surveillance Systems** – It can be integrated into smart surveillance systems to monitor and analyze human behavior in real-time

VI. DISCUSSION

Despite these promising results, there are a few concerns that remain. One of the main challenges is the detection of micro-expressions, which are very subtle and occur for a short duration. Sometimes these changes are too small to be captured accurately, which can affect the overall performance of the system.

Another issue is the variability in human behavior. A model that performs well on one dataset or group of individuals may not perform equally well on another. This lack of generalization is a common problem and highlights the need for more diverse and realistic datasets.

There is also the issue of environmental factors such as lighting conditions, camera quality, and background noise, which can affect both video and audio analysis. These variations make real-time detection more challenging.

Lastly, the computational cost of the multimodal system is relatively high, especially when processing both video and audio data simultaneously. This can impact real-time performance. However, the inclusion of multimodal analysis and confidence-based outputs is a step in the right direction. It may not solve all challenges, but it improves the reliability and interpretability of the system. The framework can be further extended and optimized for real-time applications and more advanced behavioral analysis.

VII. CONCLUSION

This project presents an AI-based multimodal framework for truth and lie detection by combining both video and audio analysis. The system integrates multiple stages such as preprocessing, feature extraction, feature fusion, and classification to effectively detect deceptive behavior.

By using CNN-LSTM for video and audio feature extraction techniques, the model captures both spatial and temporal patterns. The results show that the multimodal approach improves accuracy compared to single-modality systems. The inclusion of confidence scores also enhances interpretability.

Overall, the proposed system provides a reliable and non-intrusive solution for deception detection. In the future, it can be extended for real-time applications and improved with more advanced models and datasets.

REFERENCES

[1]. T. R. Rote, Shivama, and M. Narayanamoorthi, “Lie detection using audio classification”, Springer, 2025, doi: 10.1007/s00521-025-11485.

[2] S. Chaskar, R. Hingad, D. Muchhala, A. Dagli, and A. Kore, “Lie detection using facial expressions and speech analysis,” Springer,2025.

[3] A. Galli, M. Gravina, A. E. Pascarella, F. Di Serio, D. Cipollaro, V. Moscato, and C. Sansone, “Truth or lie: An audio-visual approach to deception detection”, Springer, 2026.

[4] T. Nagale and A. Khandare, “Comprehensive review of lie detection in subject based deceit identification,” in *Proc. Int. Conf. on Intelligent Computing and Networking (IC-ICN 2023), Lecture Notes in Networks and Systems*, Springer, 2023, pp. 89–105.

[5] H. Delmas, V. Denault, J. K. Burgoon, and N. E. Dunbar, “A review of automatic lie detection from facial features,” vol. 48, pp. 93–136, 2024.

[6] Pepino, L., Riera, P., & Ferrer, L. (2021). “Emotion Recognition from Speech Using Wav2Vec 2.0 Embeddings”. Interspeech 2021, pp. 3400–3404.

[7] Zhang, T., Deng, J., Tao, F., & Schuller, B. W. (2019). “Learning Deep Multimodal Affective Features for Spontaneous Speech. IEEE Transactions on Affective Computing”, 12(2), 509–520.

[8] Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Advances in Neural Information Processing Systems (NeurIPS), 33, 12449–12460.

[9] S. Hershey et al., “CNN architectures for large-scale audio classification,” in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 131–135.

[10] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2015.