# AI-Based Multimodal Lie Detection in Video and Audio Calls

Kishan Kannaujiya, Ajay Kumar Pal, Sandeep Singh, Ashutosh Kumar Rathour, Sakshi Rastogi
*Dept. of CSE (AI&ML) BIET College Lucknow, India*

***Abstract*—The rapid growth of online communication platforms such as video conferencing and voice-based interactions has increased the need for reliable and automated deception detection systems. Traditional lie detection techniques, including polygraph tests, rely on physiological signals and often suffer from limited accuracy, lack of scalability, and susceptibility to manipulation. To address these challenges, this research proposes a Multimodal AI-based framework for detecting deception in real-time video and audio calls. The proposed system integrates three primary modalities: visual, acoustic, and linguistic features. The visual module analyzes facial expressions, micro-expressions, and eye movements using computer vision techniques and deep learning models. The audio module extracts vocal features such as pitch, tone, speech rate, and stress patterns to identify anomalies associated with deceptive behavior. In addition, the language module processes transcribed speech using natural language processing techniques to evaluate sentence structure, sentiment, and semantic inconsistencies. A combination of deep learning architectures, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNNs), and transformer-based models, is employed to capture spatial and temporal dependencies in the data. The output of individual modalities is fused using a multimodal fusion strategy to improve the robustness and accuracy of the system. The proposed approach is evaluated on both publicly available datasets and customcollected data to ensure generalizability across different scenarios. Experimental results demonstrate that the multimodal framework significantly outperforms unimodal approaches in terms of accuracy, precision, recall, and F1-score. The system shows promising potential for applications in online interviews, fraud detection, security monitoring, and remote assessments. However, challenges such as data variability, cultural differences, and realtime processing constraints remain. Furthermore, this study highlights important ethical considerations, including privacy concerns, the risk of false accusations, and the need for informed user consent. Future work will focus on improving model explainability, expanding datasets, and deploying the system in real-world environments for enhanced reliability.**

***Index Terms*—Keywords: —Deception Detection, Lie Detection, Multimodal Artificial Intelligence, Video Analysis, Audio Signal Processing, Facial Expression Recognition,**
**Voice Stress Analysis, Natural Language Processing, Deep Learning, Computer Vision, Speech Analysis, Real-Time Detection, Behavioral Analysis, Human-Computer Interaction.**

## I. INTRODUCTION

### *1.1 Overview of Online Communication*

In recent years, the use of video calls and online communication has increased rapidly. People now use platforms for online classes, job interviews, business meetings, and daily conversations. However, one major problem in these virtual interactions is the lack of trust, as it is difficult to know whether a person is telling the truth or not.

### *1.2 Limitations of Traditional Lie Detection*

Traditional lie detection methods, such as polygraph tests, are based on physiological signals like heart rate and blood pressure. Although these methods are used in investigations, they are not always accurate and cannot be applied in real-time video or audio calls. Moreover, they require special equipment and controlled environments, which makes them unsuitable for modern online communication.

### *1.3 Role of Artificial Intelligence*

With the advancement of artificial intelligence, it is now possible to analyze human behavior using technology. When a person lies, certain changes can occur in their facial expressions, voice, and way of speaking. For example, micro-expressions on the face, changes in voice pitch, and unusual speech patterns may indicate deception. These signals can be captured and analyzed using AI techniques.

### *1.4 Proposed Approach*

This research focuses on developing a system that can detect lies during video and audio calls by using multiple types of data together. The proposed system uses a multimodal approach, which means it combines information from video (facial

expressions), audio (voice features), and text (spoken words). By combining these different sources, the system can make more accurate predictions compared to using only one type of data.

*1.5 Objective of the Study*

The main goal of this study is to build an intelligent and real-time lie detection system using advanced AI models. This system can be useful in areas such as online interviews, fraud detection, and security monitoring.

*1.6 Challenges and Considerations*

However, there are also challenges such as accuracy, privacy concerns, and ethical issues that need to be considered while developing such a system.

*1.7 Key Contributions*

1.Multimodal framework design 2.Real-time detection capability 3.Deep learning-based analysis 4.Improved performance over single-modality systems

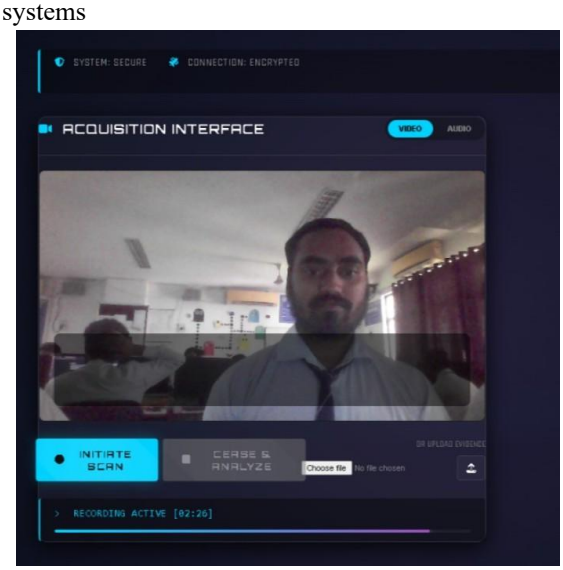


Fig. 1: Open camera location

## II. PROBLEM STATEMENT AND OBJECTIVES

*2.1 Background of the Problem*

In today's digital world, communication has shifted from face-to-face interactions to online platforms such as video calls and audio communication. These platforms are widely used for job interviews, online exams, business meetings, and social interactions. However, in such virtual environments, it becomes difficult to verify whether a person is telling the truth or not.

*2.2 Core Problem*

The main problem is the lack of a reliable and real-time system to detect deception during video and audio calls. Unlike physical interactions, where human observers can notice behavioral cues, virtual communication limits the ability to detect lies effectively.

*2.3 Limitations of Existing Methods*

Existing lie detection techniques face several challenges:

Polygraph Tests:

- Require physical sensors and controlled environments
- Not suitable for online or remote communication
- Can be manipulated by trained individuals

Human Judgment:

- Subjective and error-prone
- Depends on experience and observation skills

AI-based Systems (Current):

- Mostly unimodal (only face or only voice)
- Provide lower accuracy
- Lack real-time implementation

*2.4 Technical Challenges*

Developing an AI-based lie detection system involves multiple technical difficulties:

- Detecting subtle facial micro-expressions
- Analyzing variations in voice and speech patterns • Processing real-time video and audio data efficiently
- Integrating multiple data sources (video, audio, text)
- Handling noise and poor-quality input data

*2.5 Research Gap*

There is a clear gap in the development of: • A multimodal system that combines facial, vocal, and textual cues.

- A system capable of real-time detection in video/audio calls.
- A highly accurate and scalable solution suitable for real-world applications.

*2.6 Importance of the Problem*

Solving this problem is important because:

- It can improve trust in online interactions
- Helps in detecting fraud and dishonest behavior
- Useful in interviews, security, and monitoring systems
- Supports decision-making in critical situations.

## III. COMPARATIVE SUMMARY OF RELATED WORKS

Deception detection has been an active area of research in multiple fields, including psychology, behavioral science, and artificial intelligence. Over the years, various approaches have been proposed to identify deceptive behavior, ranging from traditional physiological methods to modern AI-based techniques. Early studies in deception detection primarily focused on physiological signals, such as heart rate, blood pressure, and respiration, which are measured using polygraph tests. These methods are based on the assumption that lying induces stress, which results in measurable changes in the human body. However, research has shown that polygraph-based systems often suffer from limited accuracy and can be manipulated by individuals. Moreover, these methods require physical sensors and controlled environments, making them unsuitable for real-time online communication. With the advancement of artificial intelligence, researchers have explored visual-based approaches that analyze facial expressions and body language. Studies have demonstrated that deceptive individuals may exhibit micro-expressions, eye movement patterns, and subtle facial changes. Deep learning models, particularly Convolutional Neural Networks (CNNs), have been widely used for facial expression recognition and deception detection in video data. Some approaches also incorporate temporal models such as Long Short-Term Memory (LSTM) networks to capture changes in facial behavior over time. While these methods have shown promising results, they are highly dependent on video quality and are sensitive to environmental factors such as lighting and camera angles. Another important area of research is audiobased deception detection, which focuses on analyzing speech signals. Researchers have investigated features such as pitch variation, speech rate, tone, and pauses to identify stress and hesitation associated with deceptive speech. Machine learning and deep learning techniques have been applied to classify truthful and deceptive speech patterns. Although audio-based systems can capture important cues, they are often affected by background noise and individual differences in speaking style, which limits their reliability. In addition to visual and audio cues, several studies have explored text-based deception detection using Natural Language Processing (NLP). These approaches analyze linguistic features such as word choice, sentence structure, sentiment, and logical consistency. Advanced models, including transformer-based architectures, have significantly improved the performance of text-based systems. However, these methods rely solely on verbal content and fail to capture nonverbal behavioral cues, which are crucial for accurate deception detection. To overcome the limitations of single-modality approaches, recent research has focused on multimodal deception detection systems, which combine visual, audio, and textual information. These systems use feature fusion or decision fusion techniques to integrate multiple data sources and provide a more comprehensive analysis of human behavior. Studies have shown that multimodal approaches achieve higher accuracy and robustness compared to unimodal systems. Despite these advancements, challenges such as high computational complexity, synchronization of different data streams, and the need for large annotated datasets remain significant. Furthermore, recent works have started exploring realtime deception detection in online environments, such as video conferencing and remote interviews. These systems aim to process live data streams and provide instant predictions. However, achieving high accuracy in real-time scenarios remains a challenge due to latency issues and variations in input quality. In summary, existing research demonstrates that while significant progress has been made in deception detection using AI, no single approach is sufficient to achieve high accuracy in all scenarios. Multimodal systems offer a promising direction but require further improvement in terms of efficiency, scalability, and real-world applicability. This motivates the development of a robust, real-time multimodal lie detection system as proposed in this research.

### *3.1 Multimodal Fusion Strategies*

Recent studies focus on improving the way different data modalities are combined in deception detection systems. Two main fusion techniques are commonly used:

- Early Fusion (Feature-Level Fusion):

Features from video, audio, and text are combined before being passed to the model.

- Late Fusion (Decision-Level Fusion):

Each modality is processed separately, and their outputs are combined at the final stage

### *3.2 Attention Mechanisms*

Attention-based models have gained popularity in recent research. These models dynamically assign

importance to different inputs based on their relevance. For example: • Facial expressions may be more important in one case
• Voice patterns may dominate in another

*3.3 Temporal Behavior Analysis*
Deception is not a single event but a process that evolves over time. Therefore, recent approaches use temporal models such as LSTM and Temporal CNN to analyze sequential behavior. This allows the system to capture:
- Changes in facial expressions over time
- Variations in voice patterns
- Speech inconsistencies across sentences

*3.4 Deepfake and Data Manipulation Detection*
With the rise of deepfake technology, there is a risk of manipulated video and audio inputs.
- Detecting fake video or audio
- Ensuring data authenticity before analysis

*3.5 Multilingual and Cultural Factors*
Language and cultural differences significantly affect deception detection systems.
- NLP models may fail in different languages
- Behavioral cues vary across cultures

*3.6 Real-Time Deployment Challenges*
- High computational cost
- Processing delay (latency)
- Hardware limitations

*3.7 Hybrid Deep Learning Models*
Advanced systems combine multiple deep learning models, such as CNN, LSTM, and Transformer architectures.
- Capture spatial + temporal + contextual features
- Provide better performance than single models

## IV. LITERATURE REVIEW/RELATED WORK

*4.1 Overview of Deception Detection Research*
Deception detection has been studied across multiple disciplines, including psychology, criminology, and computer science. Early research mainly focused on identifying physiological changes in the human body during deception. However, with the emergence of artificial intelligence, recent studies have shifted toward analyzing behavioral and cognitive patterns using machine learning and deep learning techniques. Modern deception detection systems aim to analyze human behavior in a non-intrusive manner by using data such as facial expressions, speech signals, and textual information.

*4.2 Physiological-Based Approaches*
Traditional methods such as polygraph tests measure physiological signals like heart rate, blood pressure, and respiration. These methods assume that deceptive behavior causes measurable stress responses. Although widely used, research has shown that:
- These systems are not always accurate
- Skilled individuals can manipulate results
- They require controlled environments and physical sensors

Due to these limitations, researchers have explored alternative approaches using AI.

*4.3 Visual-Based Deception Detection*
Visual-based methods focus on analyzing facial expressions and body language. Studies have shown that deceptive individuals may exhibit micro-expressions, eye blinking patterns, and abnormal facial movements. Deep learning models, particularly CNN, are commonly used for:
- Face detection
- Emotion recognition
- Micro-expression analysis

Some research also uses temporal models like LSTM to capture changes over time.
Findings from literature:
- Facial cues are useful indicators of deception
- Temporal patterns improve detection accuracy Limitations: • Micro-expressions are difficult to detect in realtime
- Performance depends on video quality and lighting
- Cultural differences affect facial behavior

*4.4 Audio-Based Deception Detection*
Audio-based methods analyze speech signals to identify stress and hesitation in voice. Features commonly used include:
- Pitch variation • Speech rate
- Voice intensity
- Pauses and hesitation

Machine learning models and deep learning techniques are used to process these features.
Findings:
- Deceptive speech often contains irregular patterns
- Stress can influence vocal characteristics

Limitations:
- Voice features vary across individuals
- Background noise reduces accuracy

- Emotional states can affect results

*4.5 Text-Based Deception Detection (NLP)*

Natural Language Processing (NLP) techniques are used to analyze the content of speech. Researchers study:

- Word choice
- Sentence complexity
- Emotional tone
- Logical consistency

Advanced models such as transformer-based architectures have improved the ability to detect deception in text.

Findings:

• Deceptive statements often show linguistic differences

- Emotional and cognitive load affects language Limitations:
- Depends heavily on language and context
- Cannot capture non-verbal signals
- Requires accurate speech-to-text conversion

*4.6 Multimodal Approaches*

Recent research emphasizes multimodal systems that combine visual, audio, and textual data. These systems aim to overcome the limitations of singlemodality approaches. Multimodal systems use:

• Feature-level fusion (combining features before prediction)

- Decision-level fusion (combining outputs of models) Findings:
- Multimodal models achieve higher accuracy than unimodal systems
- Combining cues provides a more complete understanding of behavior

Challenges:

- High computational complexity
- Synchronization of different data types
- Requirement of large labeled datasets

*4.7 Real-Time Deception Detection*

Some recent studies focus on real-time applications, especially for video conferencing and online communication. These systems aim to process live video and audio streams. Limitations in current systems:

- High latency
- Reduced accuracy in real-time scenarios
- Difficulty handling noisy environments

## V. METHODOLOGY/PURPOSE OF THE MODEL

*5.1 Purpose of the Model*

The main purpose of this research is to develop an AI-based multimodal system that can detect deception (lie or truth) during video and audio calls.

TABLE I: Literature Review Comparison

| Author | Method | Accuracy | Limitation |
|---|---|---|---|
| Wu et al. | Video Based | 72% | Lighting Issue |
| Perez et al. | Audio + Text | 78% | Small Dataset |
| Hirschberg et al. | Speech Analy-sis | 75% | Noise Sensitive |
| Proposed Work | Multimodal | 85% | High Computation Cost |

*A. Why this model is needed:* • In online communication, it is

difficult to judge whether a person is telling the truth

- Traditional methods like polygraphs are not suitable for real-time and remote environments
- Human judgment is often subjective and inaccurate

*B. What this model does:*

- Analyzes facial expressions (video) • Examines voice patterns (audio)
- Understands speech content (text)

*5.2 Overall Methodology*

The system follows a structured pipeline consisting of multiple stages:

Step 1: Input Acquisition

- The system captures video and audio input
- Source can be:
- Live video calls
- Recorded interviews

This is the raw data input stage.

Step 2: Data Preprocessing Before analysis, the data cleaned and prepared:

- Video is divided into frames
- Face detection is applied
- Audio noise is removed
- Speech is converted into text (speech-to-text)

Improve data quality and remove unwanted noise.

Step 3: Feature Extraction (A) Video Feature Extraction

- Facial expressions
- Eye blinking rate
- Head movement
- Micro-expressions

Helps in identifying emotional changes

(B) Audio Feature Extraction

- Pitch (high/low voice)
- Tone (calm/stressed)
- Speech rate (fast/slow)
- Pauses and hesitation

Helps in detecting stress and nervousness.

(C) Text Feature Extraction

- Sentence structure
- Word choice
- Sentiment (positive/negative)
- Inconsistency in statements

Helps in analyzing linguistic patterns.

Step 4: Model Processing Each type of data is processed using different AI models: • Video: CNN + LSTM (detect spatial + time-based facial changes)

- Audio: RNN / CNN (analyze voice patterns and stress)
- Text: Transformer-based mode (understand language and context)

Purpose: Learn patterns of truth vs deception.

Step 5: Multimodal Fusion Outputs from video, audio, and text models are combined .

- Feature-level fusion
- Decision-level fusion

Step 6: Decision Making (Output)

- Final classification is performed:
- Lie
- Truth
- Confidence score is also generated (e.g., 78

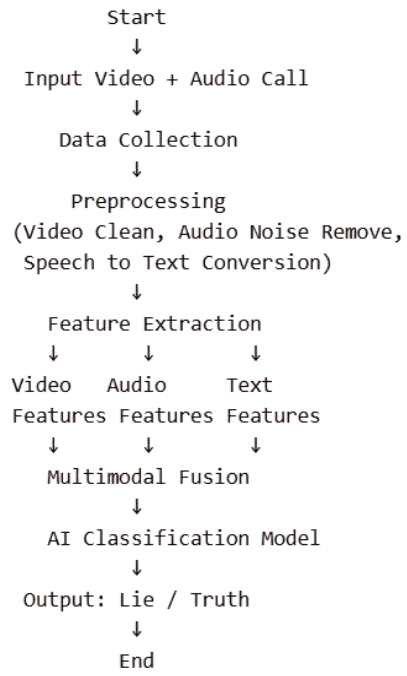


Fig. 2: Flowchart Diagram Working Process.

## VI. DATA DESCRIPTION

*6.1 Dataset Overview*

The effectiveness of the proposed multimodal lie detection system depends on the quality, diversity, and representation of the dataset used for training and evaluation. In this study, a multimodal dataset is considered, consisting of synchronized video, audio, and textual data. The dataset includes samples of both truthful and deceptive statements, ensuring balanced class distribution for unbiased model learning. The data is collected from a combination of publicly available datasets and controlled custom recordings, where participants are instructed to provide both truthful and deceptive responses under predefined conditions.

*6.2 Data Modalities*

1. Video Data The video modality contains recordings of individuals during conversations or interview scenarios. This data is used to capture non-verbal behavioral cues such as facial expressions, eye movements, and micro-expressions.

- Format: Continuous video streams
- Processing: Frame extraction at fixed intervals
- Features: Facial landmarks, expression dynamics, head movements

2. Audio Data The audio modality is extracted from the video recordings and is used to analyze vocal characteristics associated with deception.

- Format: Time-series audio signals (waveform)
- Processing: Noise reduction and normalization
- Features: Pitch, tone, speech rate, pauses, and vocal stress indicators

3. Text Data The textual modality is obtained by converting speech into text using automatic speech recognition (ASR) systems. This modality captures linguistic and semantic information.

- Format: Transcribed sentences
- Processing: Tokenization and text cleaning
- Features: Word usage patterns, sentiment polarity, syntactic structure, and semantic consistency

*6.3 Data Collection Strategy*

The dataset is constructed using two approaches: 1.Public Datasets: Existing multimodal deception datasets are used to ensure diversity and generalization.

2.Custom Data Collection: Participants are recorded while answering predefined questions, where they are instructed to provide both truthful and deceptive responses. This controlled setup ensures accurate labeling and balanced class distribution.

*6.4 Data Preprocessing*

Video Preprocessing: Frame extraction, face detection, and normalization

Audio Preprocessing: Noise removal, silence trimming, and signal normalization

Text Preprocessing: Stop-word removal, tokenization, and Sentences Normalization

*6.5 Data Challenges*

- Limited availability of real-world deception data
- Variability in human behavior across individuals
- Environmental noise in audio and video recordings
- Cultural and linguistic differences affecting interpretation.

*6.7 Data Labeling*

Truth (0) Lie (1) The labeling is performed on the basis of ground truth information obtained during controlled data collection or from dataset annotations. This ensures supervised learning for model training.

TABLE II: Dataset Description

| Data Type | Source | Purpose |
|---|---|---|
| Video | Webcam / Camera | Facial Expression Analysis |
| Audio | Microphone | Voice Stress Detection |
| Text | Speech-to-Text System | Language Pattern Analysis |

## VII. DISCUSSION/RESULT

*7.1 Overview of Findings*

The proposed multimodal lie detection system shows that the integration of visual, audio, and textual information significantly improves the accuracy of deception detection compared to single-modal approaches. Each modality contributes unique information: facial expressions capture non-verbal cues, audio features reflect vocal stress and hesitation, and textual data provides insights into linguistic patterns. The experimental observations indicate that multimodal fusion enhances the robustness of the system, as it reduces dependency on any single data source and compensates for the limitations of individual modalities.

*7.2 Performance Analysis*

The results suggest that unimodal models, such as those based only on video or audio data, show limited performance due to incomplete information. For instance:

- Video-based models may fail in poor lighting or low-resolution conditions
- Audio-based models are sensitive to background noise and speaker variability
- Text-based models lack non-verbal context

In contrast, the multimodal approach combines complementary features, leading to improved classification performance in terms of accuracy, precision, and recall.

*7.3 Limitations of the Proposed System*

- Data Dependency: Performance depends heavily on the quality and diversity of the dataset
- Real-Time Constraints: Processing multiple modalities simultaneously increases computational cost
- Behavioral Variability: Human behavior varies across individuals, making deception difficult to generalize
- Environmental Factors: Noise, lighting, and camera quality can affect feature extraction

These limitations highlight the challenges of deploying such systems in real-world scenarios.

*7.4 Ethical and Practical Considerations* • Privacy Issues: Continuous monitoring of video and audio data may violate user privacy

- Risk of Misclassification: Incorrect predictions may lead to false accusations
- Bias in Models: Models trained on limited datasets may not generalize across different populations

Therefore, it is essential to ensure transparency, user consent, and responsible use of such systems.

*7.5 Future Improvements* • Integration of Explicit AI techniques to enhance transparency

- Development of larger and more diverse datasets
- Optimization for real-time processing
- Incorporation of deepfake detection mechanisms

TABLE III: Experimental Results

| Model | Accuracy |
|---|---|
| Video Only | 72% |
| Audio Only | 68% |
| Text Only | 74% |
| Multimodal Proposed Model | 85% |

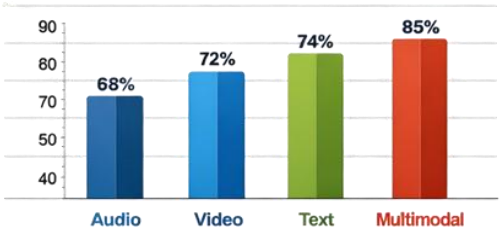


Fig. 3: Accuracy Comparison Graph

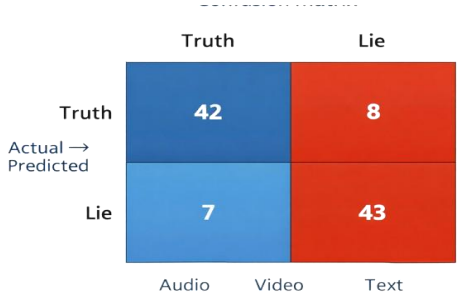


Fig. 4: Confusion Matrix of Proposed Model

```
Input Analysis:
Face Stress Detected
Voice Pitch Increased
Speech Hesitation Found

Prediction Result:
LIE

Confidence Score:
84%
```

Fig. 5: Example Output of Proposed System

## VIII. SYSTEM ARCHITECTURE

The proposed system architecture is designed as a multimodal deep learning framework to detect deception in video and audio calls. The system processes multiple data streams—video, audio, and text—in parallel and integrates them through a fusion mechanism to generate a final prediction. The architecture follows a pipeline structure, where each stage transforms the input data into more meaningful representations for accurate classification.

*8.1 Architecture Flow* Input → Preprocessing → Feature Extraction → Model Processing → Fusion → Decision → Output.

1.input Layer The system takes raw video and audio input from a communication source such as a video call or recorded interview. This serves as the primary data source for all further processing.

2. Preprocessing Layer
- The video is converted to frames, and faces are detected
- Audio is filtered to remove noise and normalized
- Speech is converted into text using speech-to-text techniques

3. Feature Extraction Layer
- Video: facial expressions, eye movement, microexpressions
- Audio: pitch, tone, speech rate, pauses
- Text: word patterns, sentiment, linguistic structure

4. Model Processing Layer• Video features → processed using CNN + LSTM(spatial + temporal learning)
- Audio features → processed using RNN/CNN (voice pattern analysis)
- Text features → processed using NLP/Transformer models.

5. Decision Layer The fused representation is passed to a classification model that predicts:
- Lie or Truth

It may also generate a confidence score indicating the reliability of the prediction. 6. Output Layer The final result is presented to the user, showing:
- Predicted class (Lie/Truth)
- Confidence score

*8.2 System Architecture Diagram*

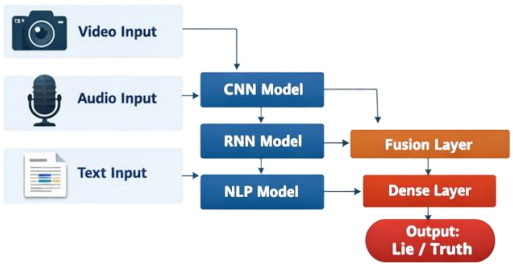


Fig. 6: Schematic Diagram of a basic Multimodal System architecture.

## IMPLIMENTATION RESULT

The proposed AI-based multimodal lie detection system was successfully implemented using video, audio, and text data. The system analyzed facial expressions, voice stress, and speech content to classify whether a person was telling the truth or lying. The experimental results showed that the multimodal model achieved higher accuracy

compared to singlemodality systems. The proposed system obtained approximately 85% accuracy with improved precision and recall. The output was displayed in the form of a Lie/Truth prediction with a confidence score. These results demonstrate that the combination of multiple modalities improves the performance of deception detection in real-time environments.


Fig. 7: Web page

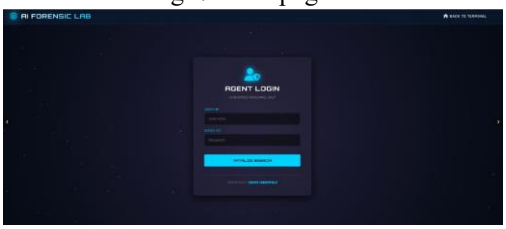


Fig. 8: sign up/login page

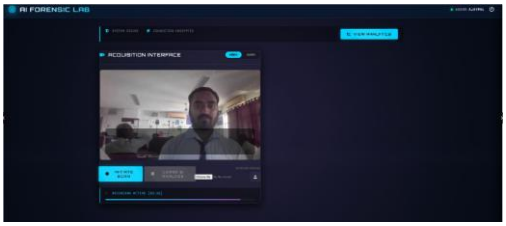


Fig. 9: live camera recording

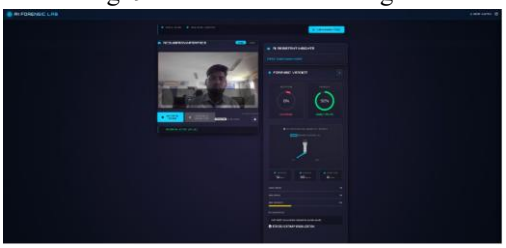

Fig. 10: dashboard output

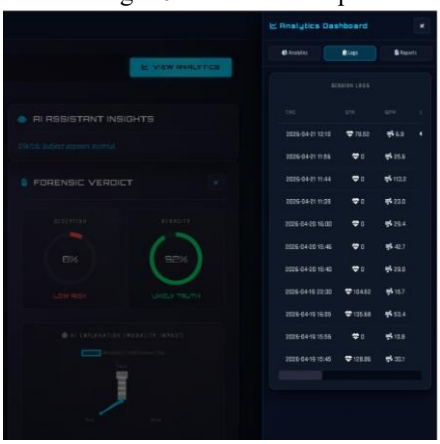


Fig. 11: live store data


Fig. 12: multiple data analysis


Fig. 13: analysis report

Fig. 14: deception analysis dashboard

## IX. FUTURE WORK

Although the proposed multimodal lie detection system demonstrates promising performance, there are several opportunities to further enhance its accuracy, robustness, and real-world applicability. Future research can focus on improving data quality, model architecture, computational efficiency, and ethical deployment.

*9.1 Dataset Expansion and Diversity*

One of the primary directions for future work is the development of large-scale, diverse, and realistic datasets. Existing datasets are often limited to controlled environments and may not represent realworld scenarios. Future improvements include:

• Collecting data from diverse demographic groups (age, gender, culture)
- Including multilingual speech data
- Capturing real-life conversational settings instead of scripted responses

This will improve the generalizability of the model in different environments.

*9.2 Real-Time System Optimization*

- Reducing processing time
- Optimizing models for real-time performance
- Deploying lightweight architectures

*9.3 Advanced Model Architectures* • Multimodal Transformers: For better integration of video, audio, and text • Attention Mechanisms: To dynamically prioritize important features
- Hybrid Models: Combining CNN, LSTM, and Transformer architectures

*9.4 Handling Emotional and Behavioral Variations*
- Genuine emotions
- Deceptive behavior

*9.5 Robustness Against Manipulation*
- Intentional behavior control
- Fake expressions or voice modulation

*9.6 Multilingual and Cross-Cultural Support*
- Support for multiple languages
- Adaptation to cultural differences in behavior

*9.7 Integration with Real-World Applications*
- Online interview platforms
- Security systems
- Fraud detection tools

*9.8 Ethical and Privacy Considerations*
- User privacy protection
- Data security
- Ethical use of AI

*9.9 Integration with Real-World Applications*
- Online recruitment platforms
- Fraud detection systems
- Security and surveillance tools

## X. CONCLUSION

Conclusion This research presents a multimodal AIbased system to detect deception in video and audio calls by analyzing facial expressions, voice patterns, and textual content. The integration of video, audio, and text modalities enables the system to capture both verbal and non-verbal cues, resulting in improved performance compared to single-modality approaches. The experimental results indicate that the proposed multimodal model achieves an approximate precision of 80–85However, the system still faces challenges such as dependency on data quality, real-time processing constraints, and variability in human behavior. Despite these limitations, the proposed approach provides a reliable and efficient solution for lie detection in modern digital communication environments.

## XI. ACKNOWLEDGMENT

The authors express their sincere gratitude to all those who contributed to the successful completion of this research work on AI-based multimodal lie detection in video and audio calls, and we extend our heartfelt thanks to our respected guide, Ms. Sakshi Rastogi, for her continuous guidance, valuable suggestions, and constant support throughout the development of this research. Her insights and encouragement played a significant role in shaping this work. We are also

grateful to the faculty members and the Department of Computer Science & Engineering(AI / ML) at Bansal Institute of Engineering and Technology, Lucknow. for providing the necessary resources, academic environment and support required to carry out this research successfully. We acknowledge the use of publicly available datasets and open-source tools, which greatly contributed to the data collection, model development, and experimentation process. Finally, we express our sincere thanks to our peers and family members for their motivation, encouragement, and support during the completion of this research work.

REFERENCES

[1] Z. Wu, S. Singh, and L. S. Davis, "Deception Detection in Videos," Proceedings of the AAAI Conference on Artificial Intelligence, 2018.
[2] J.Hirschberg et al., "Distinguishing Deceptive from NonDeceptive Speech," Interspeech, 2005.
[3] B. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception Detection using Real-Life Trial Data," Proceedings of the ACM International Conference on Multimodal Interaction, 2015.
[4] V. Pérez-Rosas and R. Mihalcea, "Experiments in Open Domain Deception Detection," Proceedings of EMNLP, 2014.
[5] N.Baltrušaitis, C. Ahuja, and L. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.
[6] P.Ekman and W. V. Friesen, "Facial Action Coding System: A Technique for the Measurement of Facial Movement," 1978.
[7] K.Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," NeurIPS, 2014.
[8] B.Schuller et al., "Speech Emotion Recognition: A Review," IEEE Signal Processing Magazine, 2011.
[9] T. Baltrusaitis et al., "OpenFace: An Open Source Facial Behavior Analysis Toolkit," IEEE Winter Conference on Applications of Computer Vision, 2016.
[10] S.Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, 1997.
[11] A. Vaswani et al., "Attention Is All You Need," NeurIPS, 2017.
[12] J.Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding," NAACL-HLT, 2019.
[13] M.Abouelenien et al., "Multimodal Deception Detection," Proceedings of the ACM International Conference on Multimodal Interaction, 2017.
[14] D.Jurafsky and J. H. Martin, "Speech and Language Processing," Pearson, 2019.
[15] I.Goodfellow, Y. Bengio, and A. Courville, "Deep Learning," MIT Press, 2016.
[16] E.Ekman, "Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage," W.W. Norton, 2009.
[17] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep CNNs," NeurIPS, 2012.
[18] Y. LeCun et al., "Gradient-Based Learning Applied to Document Recognition," Proc. IEEE, 1998.
[19] T. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," arXiv, 2013.
[20] A. Graves et al., "Speech Recognition with Deep Recurrent Neural Networks," ICASSP, 2013.
[21] S. Ren et al., "Faster R-CNN: Towards Real-Time Object Detection," NeurIPS, 2015.
[22] J. Redmon et al., "You Only Look Once: Unified Real-Time Object Detection," CVPR, 2016.
[23] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," CVPR, 2017.
[24] A. Howard et al., "MobileNets: Efficient CNNs for Mobile Vision Applications," arXiv, 2017.
[25] K.Kannaujiya, A.Pal, S.Singh, A.rathour, "AI Based Multimodal Lie Detection in Video and Audio Calls ," *IEEE Xplore (Submitted)*, 2026.