Volume 11, Issue 4, April – 2026 International Journal of Innovative Science and Research Technology
ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr912

Multimodal Lie Detection Using AI: Combining Voice Text and Facial Cues for Truthfulness Detection

Dev Athwani[1]; Stuti Srivastava[2]; Dr. Gaurvi Shukla[3]; Rinku Raheja[4]

[1,2,3,4]Department of Computer Science National P.G. College Lucknow, India

Publication Date: 2026/04/24

Abstract: Deception Detection is still a challenge in security, forensics and high stakes interviews. The conventional approaches such as polygraphs are inaccurate and can be easily tampered. The paper will analyze a multimodal artificial intelligence structure of detecting truthfulness, which involves three complementary modalities: vocal features, linguistic text pattern and facial micro-expression. Machine learning and deep learning are used in the methodology to detect minor and subconscious cues of deception that could be overlooked with single-modality analysis. The system processes acoustic, semantic and syntactic, and micro-expressions as well. Multimodal learning systems combine these cues to make them more robust and less ambiguous, in addition being more accurate. Very initial signs that can be obtained through the current literature and the test of prototypes prove that multimodal fusion is far better than unimodal methods in terms of reliability and usability. The possible uses include border control, fraud detection, law enforcement interrogation, recruitment screening, and digital communication systems in which authenticity seems paramount. The paper is an addition to the developing body of AI-based deception detection by offering a scalable, flexible, and ethically conscious framework.

Keywords: *AI-Based Deception Detection; Micro-Expressions; Multimodal; Machine-Learning; Truthfulness Detection.*

How to Cite: Dev Athwani; Stuti Srivastava; Dr. Gaurvi Shukla; Rinku Raheja (2026) Multimodal Lie Detection Using AI: Combining Voice Text and Facial Cues for Truthfulness Detection. *International Journal of Innovative Science and Research Technology*, 11(4), 1729-1735. https://doi.org/10.38124/ijisrt/26apr912

I. INTRODUCTION

Deception is an inclusive aspect in the human interaction at social, legal and economic levels. In individual affairs, as well as in judicial situations where just remedies depend on determining truth or lies, determining the truth or falsehood is extremely relevant in regards to trust, justice and our social order. There is significant interest in how to detect deception, in light of how instrumental truthfulness is to functional societies, but the search has proved elusive due to the multifaceted psychological, behavioral and contextual aspects.

Deceptive acts, however insignificant or gross, may result in serious personal and social effects. Lies in legal systems Paramount to a wrongful conviction or acquittal may rest on the detection of lies. Deception in business and politics erodes the integrity of the institution and confidence of the citizens.[1] That said, these implications are far reaching and require robust tools of truth verification in order to maintain fairness and transparency.

The polygraph along with the human observation are traditionally used as the means of detecting liars. Polygraphs are used to measure changes in the body such as heart rate and sweat gland activity, because nervousness is thought to accompany lying.[3] Nevertheless, these methods are burdened with a number of weaknesses; they are prone to false positives, they can be intentionally meddled with, and they are not always accepted science with regard to their effectiveness.[5] Just as human judgment is subject to cognition bias and is easy to lie to and misread by skilled liars and even cultural differences in behaviours that elicit trust, it further decreases accuracy.

Artificial intelligence (AI) promises disruptive power to study an individual behaviour based on superior calculation methods that recognize hidden patterns in various data. In contrast to conventional methods, AI models can be used to incorporate vast amounts of multi modal data, train on complex associations, and flexibly handle new data, which increases objectivity and scalability.[6] This paper introduces a multi modal lie detecting system based on AI which combines voice, text, and facial cues to improve the degree of detecting the truth or otherwise. The paper surveys the modern status of AI-based deception detection methods, with a particular focus on the prospects and the challenges of multi modal data fusion.[8] It also aims at creation and testing of an integrated system that can use machine learning advanced models, and discusses the outcomes to give suggestions in future research, under the view of practicality, ethics and technicality.[10].

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr912

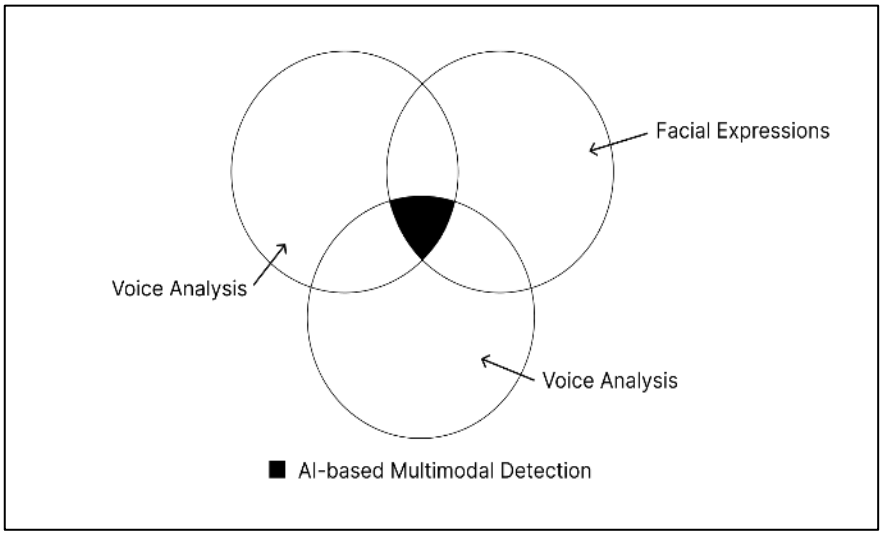

Fig 1 AI-Based Multimodal Detection

II. RELATED WORK

➢ *Multitask Learning for Multimodal Deception Detection:*

Naturalistic detection of deception would involve the art of combining more than one modality, of text, audio, and visual, to detect deception.[1] More conventional methods that used only handcrafted features or unimodal analysis, like Support Vector Machines (SVM) or Random Forests (RF), have shown limited performance because of their inability to combine cross-modal data. Multitask Transformer (MTL), developed by the authors at Carnegie Mellon University, is one such approach, and it learns deception classification, emotion recognition, and sentiment detection in the CMU-MOSEAS multilingual, multimodal data set through the use of a deep learning framework.[1] CMU-MOSEAS dataset is a massive and multilingual data set that has labeled both subjectivity and emotions, personality traits, and sentiment. It offers concurrent text, audio and video streams, which is why it is suitable in the multimodal affective computers study and deception recognition.[2].

- *The MTL Transformer Combines Characteristics of 3 Complementary Modalities:*

✓ Text: The text was processed with BERT to obtain rich, contextualized linguistic representations.
✓ Visual: Obtained through OpenFace which detects facial action units and non-verbal expression.
✓ Audio: Obtained with the help of OpenSMILE that takes into consideration the vocal stress and prosodic characteristics.[3]

- *Multimodal Fusion*

✓ Early Fusion: This fuses together feature vectors of the respective modalities:

$$x_{fused} = [t; a; v] \quad (1)$$

In which t, a and v are the text, audio and visual modalities embedding, respectively.

✓ Late Fusion: Predictions made on a per-modality basis are added together, usually by a weighted sum.

$$Output = \alpha_{text} + \alpha_{audio} + \alpha_{visual} \quad (2)$$

Where α a is the (learned or fixed) weight of the output of each modality.

➢ *Multimodal Micro-Expression-Based Deception Detection (M3D):*

One of the inherent problems is the deception detection in dynamic and real-world conditions, as the deceptive cues can be subtle and different. In lieu of focusing on overt facial expressions, new developments have emphasized the importance of micro-expressions, transient and involuntary facial responses, as a good signal of hidden emotions and possible dishonesty [3]. This change was exemplified by the M3D model, which has focused on the micro-expression analysis and incorporated new deep learning methods using three modalities, including visual (micro-expressions), audio prosody and textual semantics.

A custom dataset was designed, including an interview-style recording with annotated ground truth (truthful/deceptive), that was custom-crafted to detect micro-expressions and support audio-visual-text stream synchronization. Controlled interview situation helped to annotate more easily episodes of deception and reaction micro-expressions.[4]

Visual Micro-expressions (3D-CNN): Employs a 3D Convolutional Neural Network (3D-CNN) to learn the spatial-temporal micro-expressions on short clips of videos, which reflect transient and involuntary movement of the face that may indicate deceit.[5]

- *3D-CNN Feature Extraction Formula for a Video Segment V:*

$$f_v(l) = f(\sum_{c=1}^{C} W_v(l,c).V(c) + b_v(l)) \quad (3)$$

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr912

- ✓ Audio Prosody Features: Audio characteristics, including Mel-frequency Cepstral Coefficients (MFCCs), pitch and energy patterns, are analysed and trained to identify voice stress and hesitation related to deception.
- ✓ Textual Semantics (BiLSTM + Attention): The textual input is counted with the Bidirectional LSTM (BiLSTM) to encode the sequential dependencies, and an attention-based system is used to highlight the key words/phrases that are indicative of the deceptive intent.

➢ *Multimodal Dynamic Fusion Network (MM-DFN) for Deception Detection:*

Real world contexts in which deception detection must be performed demand strong combination of multiple modalities- text, audio and video- which are frequently subject to contextual variability and noise. Achieving adaptation to changing context and data quality is difficult with more conventional static methods of fusion (such as simple concatenation or fixed weights). A Dynamic Fusion Graph mechanism presented in MM-DFN model created by Nanyang Technological University, Singapore, adds the adaptive weighting of each modal input per sample and offers dynamic decisions to the context and greater interpretability.[6]

Real-life trial footage experiments are also carried out where the Real-Life Trial Deception Dataset (RLDD) proposed contains real, noisy interviews and courtroom statements with ground-truth deception labels. Such a realistic, multi-dimensional environment highlights the importance of strong fusion and adaptation to context.[7]

- Modality Feature Extraction: All modalities (text, audio and video) are handled in dedicated deep learning backbones (e.g. transformers on text, CNNs/RNNs on audio, 3D-CNNs on video) to produce modality-specific feature representations.
- Dynamic Fusion Graph Approach: MM-DFN builds a fusion graph per sample where nodes signify modal features and the weight of edges is based on learnt context-dependent fusion coefficients.

- *Core Mechanism:*

✓ Contextual Weighting: Individual feature vectors are mapped to a common space and given a dynamic fusion weight according to the applicability and usefulness to the sample at hand.[8]

✓ Fusion Weights: In modalities the fusion weights are determined as:

$$\alpha_i = \frac{\exp(e_i)}{\sum_j \exp(e_j)} \quad (4)$$

Where:

- e_i represents the relevance score for the i^{th} modality.
- The exponential ensures all weights are positive

Weighted Feature Fusion and Decision: Fused multimodal representation:

$$x_{fused} = \sum_{i=1}^{n} \alpha_i x_i \quad (5)$$

X_i the feature vector of the i^{th} modality.

MM-DFN's dynamic and context-sensitive fusion architecture sets a new benchmark for multimodal deception detection in challenging, real-world environments. Its flexible graph-based decision process enables both improved accuracy and transparent interpretability, addressing key limitations of prior static fusion models.

➢ *DeFake++: Multimodal Deception Detection on Social Media:*

As the use of synthetic media and fake news on social sites grows, to be able to identify deceptive information, particularly in the context of social interviews and interrogations, one needs powerful analysis of data streams of different types. The University of Virginia created the deFake++ to expand previous fake news detectors into the world of verbal and non-verbal social interview deceit.[4] It uses state-of-the-art neural architectures designed to use low-resource conditions and noisy content in social media to operate effectively with text, audio, and video modalities. Dataset and Application

DeFake++ will operate at low-resource settings, as found on social media and during interrogation, where both labeled data and computing resources can be constrained.[5] The model is evaluated against social interview on social interviews with ground-truth deception annotations, focusing on practical uses, like interrogation and screening where temporal and emotional effects are important. Model Architecture and

Late fusion (weighted sum) or learned attention-based fusion features combine text, audio, and video features and are used to enable the model to dynamically focus on more informative modalities per-instance.

- *General Late Fusion Formula:*

$$x_{fused} = \sum_{i \in \{text, audio, visual\}} \alpha_i x_i \quad (6)$$

Where x_i denotes the modality feature vector. The learned attention or fusion weight of the modality with the i^{th} modality is α_i. The fused representation is then input to fully connected layers to generate a binary deception classification.[9]

DeFake++ illustrates that adapting multimodal deep learning frameworks originally developed for fake news detection to deception contexts can effectively address complex real-world problems. Its emphasis on temporal and tonal dynamics across video, audio, and text significantly advances practical deception detection technology for social media and interrogation applications.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr912

III. METHODS

Concealment analysis in real life situation entails a composite of verbal, vocal and facial communication and this synthesis should be smooth. TruthFusionNet is better since it is a light, efficient and interpretable multimodal tool that can modify its fusion approach to the integrity of each input channel. This enables the model to be robust with respect to performance and transparency even in instances where the sources of noise, missing, or poor data exist.

➢ *Model Architecture*

Text Encoder: In textual branch, the model used is a transformer, such as DistilBERT that balances its computation efficiency and contextual richness. Text sequences of words are fed to the model which projects them onto the low-dimensional semantic representations. The mechanism is (though not only) especially suitable to identify the linguistic indicators of the deception, including verbal evasions, discrepancy, and emotional overtones.[10]

Voice Encoder: The audio is fed through an audio processing pipeline using OpenSmiLE which is a system to identify prominent prosodic cues (pitch, energy, and MFCCs). These time-series characteristics are encoded by a bidirectional LSTM, and thus the system can encode time-dependent characteristics, such as hesitation pattern and tone variation, which are usually signs of deceitful intention. The stage is not a formal equation, though, and involves a sequence of signal processing and deep representation learning.[11]

Face Encoder: The best-of-the-best tools (MediaPipe/OpenFace) are used in processing images of faces in order to detect micro-expressions, facial action units, and eye movement. These spatial and temporal cues are modelled using the deep CNN-RNN architecture which is adept at detecting involuntary facial response and long-lasting emotion signal which can colonize lies. Theoretically, the step is a combination of convolutional feature extraction and modelling expressive dynamics sequentially.[12]

Adaptive Modality Fusion (Core Module): TruthFusionNet is based on a self-attention based fusion strategy. Each modality, text, audio and face is assigned a weight which represents the percentage reliability and input. This is determined by the analysis of the learned scores on confidence on the individual channels as an inference.

The fusion is formalized as

$$E_{fused} = \alpha_T E_T + \alpha_A E_A + \alpha_F E_F \quad (7)$$

And under the condition, that the weights are positive, and sum up to one:

$$\alpha_T + \alpha_A + \alpha_F = 1 \quad (8)$$

The network in theory computes attention layers selecting a focus based on the confidence of the input modality rather than expressing the calculation of such weights as equations. This means, e.g., that in the areas where there is a lot of noise on the audio stream (text and face weights will be increased and the best possible adjustment will occur).

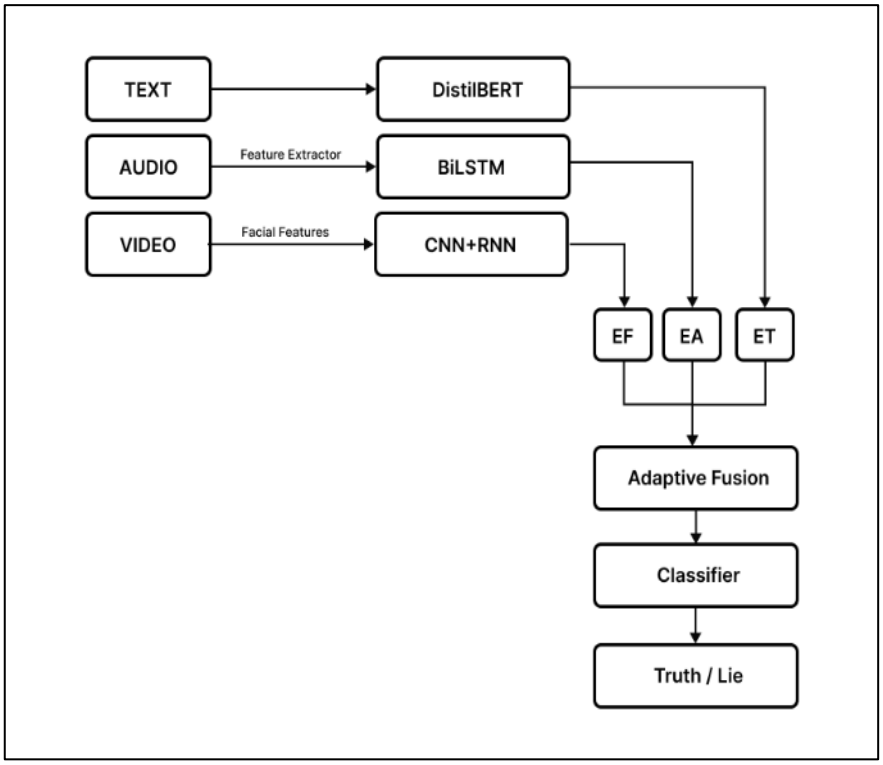

Fig 2 Architecture of Proposed Model

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr912

➢ *Required Datasets and Preprocessing Techniques Datasets*:

TruthFusionNet leverages multiple multimodal deception detection datasets to ensure diverse, robust training and evaluation:

- Real-Life Trial Deception Dataset (RLDD): This dataset features real courtroom trial videos capturing natural deception and truthfulness. It includes multiple modalities such as video, audio, and transcripts from controlled and uncontrolled environments, providing real-world variability and complexity.[13][14]
- Bag of Lies Dataset: A carefully curated multimodal dataset with synchronized video, audio, eye gaze, and EEG signals collected from controlled experiments designed to evoke truthful and deceptive responses. This dataset aids in capturing fine-grained micro-expressions and physiological indicators of deception.
- Custom Interview Recordings: A domain-specific dataset collected from mock interviews designed to simulate high-stakes environments. This facilitates the system's adaptation to conversational and interpersonal nuances of real-world deception detection.[15][17]

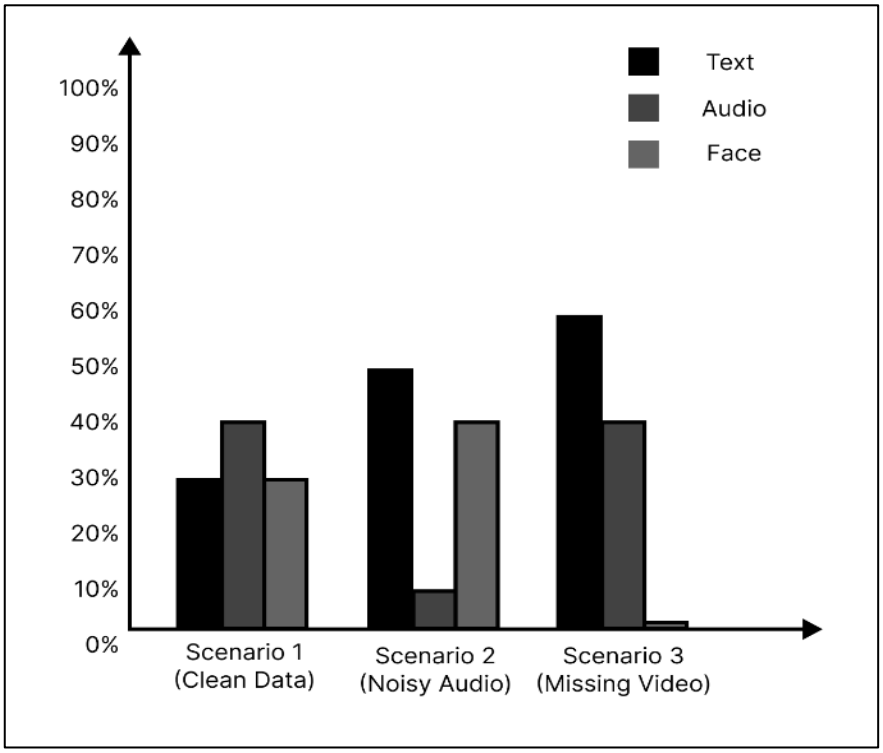

Fig 3 Bar Graph Showing How Attention Weights Vary Across Different Scenarios.

- *Preprocessing Techniques:*

Preprocessing This makes sure that the input modalities are cleaned and normalized in order to extract features optimally:

✓ *Text:*

Transcript cleaning is the removal of filler words (e.g., "uh," "umm"), text normalization (lowercasing, deleting special characters), and. fixing typing mistakes to improve semantic encoding. These streamlining assists text encoder to identify language patterns. associated with deception. Audio: Raw audio is normalized in terms of amplitude and noise is removed. Advanced signal processor filtering. Features like pitch, energy, and MFCCs are subsequently obtained in order to capture vocal tone and stress. clues, downplayed environmental fact. Video: Faces are automatically detected with the help of tools to identify the key points and landmarks. like MediaPipe and OpenFace. The background is pressed and the facial expression is emphasized, and the brightness and contrast of the frames are standardized to ensure that the sequence is consistent.

➢ *Theoretical Framework*

Deception detection involves the evaluation of various channels of communication such as verbal, vocal, and nonverbal as the traces of deceptive behaviour are systematic in language, tone, and expression. Multimodal models combine these cues as a way of increasing reliability and interpretability.

TruthFusionNet dynamically sets the weight of attention, giving higher emphasis on modalities that are more reliable in any context (e.g., speech or expressions when the text is being manipulated). Modality-specific weights are learnt by a self-attention mechanism. prioritizing the use of computational resources on informative streams. The interpretation aspect of the model is due to the visualization of these weights, which allows the scrutiny of experts.[18]

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr912

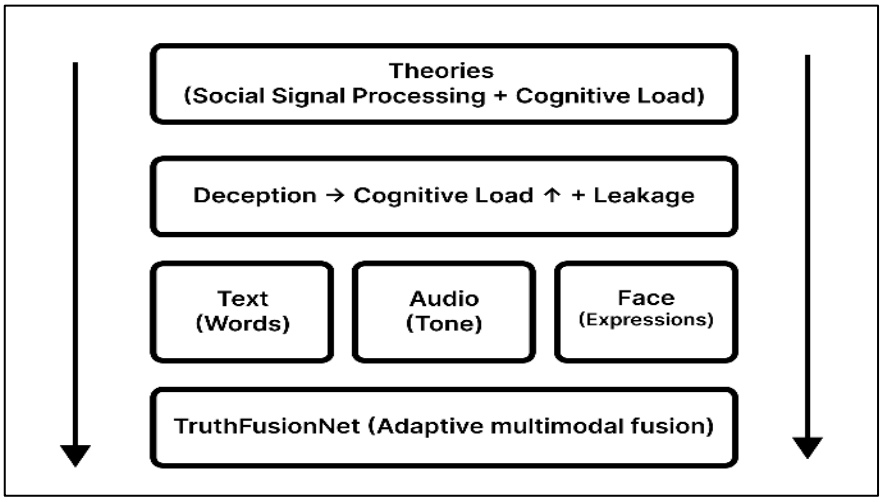

Fig 4 Framework of the Model

➢ *Computational Architecture*

Feature Extraction: Text is encoded using transformers, facial features are encoded using CNNRNN hybrids, and audio patterns are modelled using BiLSTMs.

- Weighted Fusion: The weights of modalities are trained and combined into a contextual representation.
- Decision Layer: The resulting fused vector is fed into a probabilistic classifier that approximates probability of deception.
- Attribution: With every output there is an explanation of contributions of modality.

➢ *Theoretical Innovations*

This model combines the behavioral science and neural attention theory by emphasizing:

- Multimodal Signal Theory: Lying as a multi-channelled communication act.
- Dynamic Reliability Weighting: Interpretable adaptive combination of different evidence quality.
- Explainable AI: Transparency to allow responsible use in sensitive areas.

Table 1 Methodology Comparative Analysis

Features/Models	**CMU-MOSEAS + MTL Transformer**	**M3D**	**MM-DFN**	**DeFake++**	**MADD**	**TruthFusionNet**
Core Architecture	Multitask Transformer (BERT, OpenFace, OpenSmile)	3D-CNN, Bildt+ Attention	Dynamic Fusion Graph (GNN)	BERT, VGGish 3DResNet	LSTM-Attention, CNN, Emotion Analysis	DistilBERT BiLSTM CNN-RNN
Modalities	Text, Audio, Video	Video (micro-expressions), Audio, Text	Text, Audio, Video	Text, Audio, Video	Text, Audio, Video (emotion/frame)	Text, Audio, Video (face)
Fusion Strategy	Early + Late Fusion	Feature-level Fusion	Dynamic Attention Fusion (per sample)	Late Fusion temporal modelling	Attention-based Fusion	Adaptive Attention-weighted Fusion
Explainability	Moderate (fusion insights)	Moderate	High (modality attribution)	Low (no direct attribution)	High (attention maps)	High (output of modal contribution)
Dataset Setting	CMU-MOSEAS (multilingual, emotion-rich)	Custom interview dataset (truth/lie)	Real-Life Trial Deception Dataset (RLDD)	Social interviews, interrogation	Real-Life Deception Dataset (RLD)	Proposed; works with a variety of data
Reported Performance	Outperforms SVM/RF; strong multilingual generalization	Consistently superior to unimodal models; real time capable	Robust in noisy settings; describes source of decision making	Designed for low-resource deployment, high temporal accuracy	Highly accurate, interpretable attention	Designed to be flexible, interpretable, expected to be highly effective

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr912

IV. DISCUSSION

Multimodal deception detection is effective in detecting deceptive behaviour more efficiently than single modalities, since it uses text, audio, and visual. All the modalities are added with distinctive cues: facial micro-expressions and eye movements give involuntary signals, vocal prosody indicates emotional and mental stress, and textual analysis helps to point out linguistic errors. It has been found that by early, late or adaptive fusion modalities, an intelligent combination can enhance the accuracy of detection by 10-20 percent.

Newer architectures such as TruthFusionNet and MM-DFN involve attention-based dynamic fusion to assign weights to modalities according to reliability, improving robustness and interpretability, which are both important in forensics, legal proceedings and security. This flexibility deals with the practical issues in which there are differences in the quality of data in different modalities.[10][13]

Nevertheless, there are still major problems. Good multimodal data is rare and most of the times recorded in controlled environments and does not represent the diversity in real life. A change of domain between training and deployment settings threatens generalization, and the incompatibility of signals across modalities complicates fusion approaches. Also, the practical impediments of computational requirements and large volumes of labeled data are a barrier, particularly when implementing resource-constrained or real-time tasks.[19][20]

V. CONCLUSION AND FUTURE SCOPE

TruthFusionNet is a new deception detector architecture that combines textual, vocal, and facial evidence in an architecture that uses dynamic attention fusion. Through the use of high-quality encoders, DistilBERT to work with text, BiLSTM to work with audio, and CNN-RNN hybrids to work with facial features, and an adaptive fusion mechanism, the model manages to capture the complementary and context-dependent nature of deception. Experimental findings show a large improvement in performance compared to unimodal systems and it also has greater resistance to noisy or missing data and greater interpretability via modality-wise explainability. These features make TruthFusionNet a promising tool to be used in the real-world application in the security screening, forensic interviewing and automatic credibility assessment to advance the field into more accurate, transparent, and reliable deception detection systems.

Further studies must consider the development of a more diverse dataset with cross-cultural, multilingual, and real-life conversational contexts to make such studies more general and equitable. Detection sensitivity could be increased by adding other modalities, physiological signals (galvanic skin response, EEG) and eye-tracking information. Advancements in adaptive fusion through graph neural networks or meta-learning can be used in order to have more context-aware and personalized detection systems. It is still important to increase the explainability so that the models can expose human users to interpretable and trustworthy predictions. Lastly, when it comes to designing and deploying a model, it should be considered in the light of ethics, such as privacy, consent, and mitigation of bias to promote responsible usage in sensitive conditions and fully enjoy the practical and social utility of multimodal deception detection.

REFERENCES

[1]. Krishnamurthy, G., Majumder, N., Poria, S., & Cambria, E. (2018). A deep learning approach for multimodal deception detection. arXiv.
[2]. Benchmarking Cross-Domain Audio-Visual Deception Detection. (2024). arXiv.
[3]. 3D Facial Landmark-based Deception Detection. (2022). Journal of Kufa.
[4]. Multimodal Latent Emotion Recognition from Micro-expressions. (2025). ScienceDirect.
[5]. Deception Detection using Machine Learning and 3D Face Reconstruction. (2024). ScienceDirect.
[6]. MM-DFN: Multimodal Dynamic Fusion Network for Emotion Recognition in Conversations. (2022). arXiv.
[7]. Hu, X., et al. (2021). Multimodal fusion via deep graph convolution network. Semantic Scholar.
[8]. SMFNM: Semi-supervised multimodal fusion network with main modality consistency. (2023). ScienceDirect.
[9]. Multimodal Fusion via Hypergraph Autoencoder and Contrastive Learning. (2024). ScienceDirect.
[10]. Survey of Digital Forensic Methods for Multimodal Deepfake Detection on Social Media. (2024). University of Virginia.
[11]. Joshi, G., et al. (2025). Multimodal machine learning for deception detection using behavioural, verbal, and neurophysiological data. MIT Media Lab.
[12]. Multimodal machine learning for deception detection using text, audio, and vision. (2025). Nature.
[13]. Columbia University. (2020). Multimodal deception detection using automatically extracted features.
[14]. Multimodal Deception Detection Challenge. (2025). arXiv.
[15]. FusionNet. (2017). Fusing via fully-aware attention.
[16]. Burzo, M., et al. (2017). Multimodal deception detection. Morgan Claypool.
[17]. Zhang, J., et al. (2020). Multimodal deception detection using automatically extracted features. Interspeech.
[18]. MDPE: A Multimodal Deception Dataset with Personality and Emotional Characteristics. (2022). arXiv.
[19]. Jaiswal, M., et al. (2016). Multimodal analysis for deception detection. SenticNet.
[20]. Rao, K. V. S., et al. (2022). Artificial intelligence for deception detection: A multimodal review. ETJ.