IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

DOI: 10.17148/IARJSET.2026.131169

Accuracy-Weighted Late Fusion of Visual, Acoustic, and Linguistic Features for Deception Detection

Maheta Hardikkumar Harishbhai[1], Dineshkumar Bhagwandas Vaghela[2*]

Research Scholar, Gujarat Technological University, Ahmedabad, Gujarat, India[1]

Assistant Professor, Shantilal Shah Engineering College, Bhavnagar, Gujarat, India[2*]

* Corresponding author's Email: dineshvaghela28@gmail.com

Abstract: Automated deception detection remains difficult because facial appearance, speech patterns, and language content vary across speakers and recording conditions. This study presents a decision-level framework that combines visual, acoustic, and linguistic evidence. The visual branch detects faces in sampled video frames, extracts SE-ResNet50 representations, and uses a Random Forest classifier. The acoustic branch joins eGeMAPSv02 and EmoBase descriptors from openSMILE and applies a multilayer perceptron. The linguistic branch uses RoBERTa-based representations with parameter-efficient adaptation and a Random Forest classifier. The framework combines the three probability outputs through weights proportional to the individual model accuracies. Experiments use the Real-Life Trial dataset containing 121 recordings, including 60 truthful and 61 deceptive clips. The strongest standalone audio, visual, and text configurations obtain 81.03%, 76.7%, and 71.1% accuracy, respectively. The weighted fusion obtains 86.78% accuracy, 0.9008 AUC, and 0.8678 macro F1. The study also examines how fusion changes individual predictions. Results support the usefulness of complementary evidence, while the small dataset and unresolved evaluation details limit claims about generalization.

Keywords: Deception Detection, Multimodal Learning, Late Fusion, openSMILE, SE-ResNet50, RoBERTa.

I. INTRODUCTION

Deception detection seeks to identify whether a statement is truthful or intentionally misleading. Human observers often struggle to make this judgment consistently. A speaker may change facial expression, voice quality, or word choice for reasons unrelated to deception. Therefore, a single behavioural cue cannot provide a reliable answer in every setting [10], [11], [21]. Machine learning provides a way to study several types of evidence together. Courtroom video contains a visual stream, a speech signal, and linguistic content. Each stream describes a different aspect of the same statement. Previous work has explored these signals separately and in multimodal systems [1], [3], [12], [20]. This paper studies a simple and interpretable combination of three independently trained classifiers. Instead of concatenating large feature vectors, the framework combines their output probabilities. It assigns larger weights to models with higher reported accuracy. The main question is whether this weighted combination improves the result over each standalone model on the same dataset. Deception is a complex communication behaviour rather than a single observable facial expression or speech event. A person may conceal information, provide an inaccurate account, or intentionally change the meaning of a statement. The same outward behaviour can also arise from anxiety, fatigue, unfamiliar surroundings, or the pressure of being recorded. For this reason, automated deception detection must examine patterns of evidence instead of interpreting one movement or vocal change as definitive proof of dishonesty [10], [11]. The problem is especially relevant to recorded interviews and courtroom proceedings. These settings generate synchronized visual, acoustic, and linguistic information, but they also introduce difficult recording conditions. A speaker may turn away from the camera, speak over background noise, or produce a short answer that provides little linguistic context. A practical computational framework should tolerate such differences and should avoid assuming that one modality is equally informative for every recording [1], [3]. Visual observations describe appearance and facial behaviour. Facial muscle activity and expression can be represented using engineered descriptors or deep convolutional embeddings. Nevertheless, expressions do not have a fixed one-to-one relationship with truthfulness. Lighting, pose, camera resolution, and identity may influence an image model. Audio analysis provides another view through pitch, loudness, spectral distribution, and voice quality, while linguistic analysis examines contextual meaning and word usage [13], [17], [20]. Each modality has complementary strengths and weaknesses. A visual classifier may obtain useful facial representations when the speaker is clearly visible but become less reliable when face detection fails. Acoustic features can capture changes in vocal delivery even when the face is obscured, although microphone characteristics and emotional arousal can interfere. Textual representations

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

DOI: 10.17148/IARJSET.2026.131169

can preserve semantic information despite poor image quality, but transcripts omit some nonverbal signals and may contain transcription errors. These observations motivate a combined approach [12], [14], [20].

Sehrawat et al. investigated a multimodal stacked Bi-LSTM approach [8], illustrating another strategy for learning relationships across observations. Sequence-based methods may preserve changes over time, whereas recording-level acoustic functionals summarize the entire speech segment. The present approach uses the latter representation, which makes it possible to compare conventional classifiers on a fixed-length feature vector. Derakhshan et al. examined the selection of useful features for multimodal deception detection [14]. Feature relevance is important because combining more descriptors does not automatically improve performance. Redundant features, unequal scales, and small training sets can make large concatenated representations difficult to model. Decision-level fusion reduces this dimensionality problem by operating on a small number of probability outputs. The literature indicates three recurring challenges. First, behavioural cues are ambiguous and affected by context. Second, feature extractors trained on general image, speech, or language tasks may not transfer perfectly to deception detection. Third, evaluation can be optimistic when samples from the same speaker occur in training and testing or when fusion parameters are tuned using held-out labels. These concerns motivate careful interpretation of the experimental results. The proposed study contributes to this literature through a simple accuracy-proportional weighting strategy. Rather than claiming a new feature extractor, it combines established modality-specific representations and evaluates how the resulting decision differs from the strongest standalone model. The error-correction analysis is especially useful because it separates rescued predictions from newly introduced mistakes. An important design distinction is whether a visual model retains temporal ordering. The visual pipeline averages the embeddings of sampled frames, producing a compact video-level vector. This is computationally convenient for a small dataset but does not explicitly model movement sequences. Thus, the selected visual branch should be understood as pooled facial appearance representation rather than a temporal action-recognition network. Previous studies have considered both feature-level and decision-level multimodal fusion. Feature-level approaches place different representations into a shared learning model, which can capture interactions but may require carefully aligned features and sufficient training examples. Decision-level approaches train the branches separately and then combine their predicted class probabilities. The latter design is particularly suitable when visual, acoustic, and linguistic representations differ substantially in dimensionality and model architecture [2], [3], [20]. The input streams are processed separately because each requires a different preparation stage. Video frames are sampled and filtered for valid faces, audio recordings are transformed into statistical speech descriptors, and transcripts are tokenized for contextual encoding. Keeping the branches independent allows the classifier choice to reflect the characteristics of its input representation. The fusion stage operates only after each branch has produced a comparable probability for the same recording. Probability alignment is a necessary requirement of decision-level fusion. A numerical probability is meaningful only when the three branches use the same class interpretation and refer to the same sample. In particular, an incorrectly ordered probability column could reverse one modality's contribution without changing the dimensionality of the fusion input. Consistent class-label mapping must be confirmed across the modality outputs before replication. Unlike attention-based fusion, accuracy-weighted averaging does not learn sample-dependent interactions. The weights are fixed for all recordings, and a strong modality receives a larger contribution based on its aggregate reported performance. This makes the decision easy to interpret but can be restrictive when recording quality varies substantially. The method is therefore best understood as a transparent integration rule rather than a mechanism that automatically measures the reliability of each modality for an individual speaker.

The present work investigates an accuracy-weighted decision-level framework. The visual branch uses sampled face images, an SE-ResNet50 feature extractor, and Random Forest classification. The acoustic branch combines eGeMAPSv02 and EmoBase functionals and trains an MLP. The linguistic branch uses RoBERTa representations with selective adaptation and Random Forest classification. Their outputs are combined using nonnegative weights derived from the reported standalone accuracies. The Real-Life Trial dataset provides a common experimental setting with 121 recordings, comprising 61 deceptive and 60 truthful clips [1]. Its relatively small size makes model selection and performance reporting important. The study therefore focuses on the reported comparison of individual modalities and fusion alternatives, while recognizing that results on a limited benchmark do not establish generalization to unseen speakers or recording environments. The central contribution is a transparent fusion mechanism that does not require joint end-to-end optimization of all three branches. Its weights can be inspected directly, and the fused score can be traced back to the three modality probabilities. The analysis also considers the errors corrected by fusion and the cases in which fusion introduces an error. This distinction is important because higher overall accuracy does not imply that every individual prediction improves. The remainder of the paper reviews relevant behavioural and computational research, explains the three processing pipelines and the weighted combination rule, presents the experimental results reported in this study, and discusses the limitations of the evaluation and opportunities for future work.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

DOI: 10.17148/IARJSET.2026.131169

II. Related Work

A. *Behavioural and Visual Analysis*

Early deception studies examined emotional expression and facial muscle activity. Ekman and Friesen introduced the Facial Action Coding System to describe visible facial movements [13]. Later research used computational methods to extract visual information from interviews and trial recordings. Pérez-Rosas et al. introduced the Real-Life Trial benchmark and studied multimodal deception cues [1]. Yildirim et al. examined micro-expressions for deception detection [9]. Deep convolutional networks provide a compact way to represent facial appearance. ResNet uses residual connections to train deep feature extractors [15]. Squeeze-and-Excitation networks reweight channels so that informative responses receive more attention [16]. These ideas motivate the visual branch used in this study. However, facial expressions are not direct proof of deception. Differences in lighting, identity, pose, and camera quality can also affect predictions. Behavioural research provided the conceptual foundation for computational deception analysis. Ekman and Friesen developed the Facial Action Coding System to describe facial movements systematically [13]. Related psychological work discussed emotional leakage, deliberate control of expression, and the difficulty of identifying lies from isolated cues [10], [11]. These studies motivate facial measurement, but they do not justify treating a particular expression as a universal deception marker. Pérez-Rosas et al. introduced a real-life trial benchmark that supports analysis of spontaneous statements rather than only staged interactions [1]. This setting is important because natural behaviour differs from acted deception, and the recordings contain multiple forms of information. The benchmark also creates challenges associated with a modest sample size, variable capture quality, and the possibility that speaker characteristics influence model performance. Yildirim et al. examined the role of micro-expressions in deception detection [9]. Their work reflects interest in short and subtle facial changes that can be difficult to observe manually. However, approaches relying on fine facial details remain sensitive to frame quality, face visibility, and the temporal resolution of the recording. This motivates the use of robust visual representations rather than a dependence on a single manually defined facial cue.

Deep image representations offer a different route to visual analysis. He et al. introduced residual learning to support deeper convolutional networks [15]. Hu et al. proposed Squeeze-and-Excitation mechanisms that adaptively emphasize feature channels [16]. The visual pipeline combines these ideas in its SE-ResNet50-based feature extraction stage and summarizes frame embeddings to produce one representation per recording. Visual analysis in recorded testimony is affected by the camera viewpoint, lighting, head movement, and the duration of visible facial behaviour. A face-based model may therefore learn appearance differences that do not represent deception. Earlier behavioural work provides a reason to study facial movements, but it does not establish a universal expression that identifies a false statement [10], [11], [13]. The Real-Life Trial dataset introduced by Pérez-Rosas et al. is important because it contains recordings of statements made in realistic legal settings [1]. Such data differs from acted deception collected under laboratory instructions. It also creates practical challenges, including variable recording quality and a limited number of speakers. These conditions motivate consistent preprocessing and careful evaluation.

B. *Acoustic and Linguistic Analysis*

Speech carries information through pitch, intensity, rhythm, and spectral characteristics. The openSMILE toolkit extracts standard acoustic descriptors from recordings [17]. Prior deception studies have used speech features and sequential models to represent changes in vocal behaviour [8], [18], [23]. Acoustic features may also capture stress, emotion, or recording noise rather than deception alone. The uniform sampling stage spreads selected observations across the recording rather than selecting adjacent frames. This provides broader temporal coverage at limited computational cost. It does not, however, guarantee that the sampled frames contain informative facial behaviour. If the face is not detected at a selected instant, that observation contributes no embedding to the video-level mean. The resulting representation therefore depends on both the frame-selection procedure and the success of face localization. The 2048-dimensional pooled representation is supplied to a Random Forest rather than training the entire convolutional backbone on the small benchmark. This separates generic image representation from task-specific classification and reduces the number of parameters that must be estimated from the deception labels. A limitation is that the representation may emphasize appearance characteristics unrelated to deception. Subject-independent evaluation would be valuable for measuring how much the visual model relies on transferable behavioural cues. Acoustic deception analysis examines characteristics of speech production that are not directly represented in written transcripts. Fundamental frequency, intensity, spectral energy, cepstral coefficients, and voice-quality descriptors can change across utterances. These changes may relate to emotional state, speaking style, or recording conditions as well as deception, so acoustic classification is best interpreted as pattern recognition rather than direct physiological measurement [17], [18]. The openSMILE framework provides standardized extraction of speech descriptors and summary functionals [17]. The framework uses eGeMAPSv02 and EmoBase together, joining an 88-dimensional vector with 988 functionals. This combination brings together a compact, interpretable feature family and a broader set of low-level statistical descriptors. The resulting 1076-dimensional vector is supplied to conventional classifiers, with the MLP giving the strongest reported acoustic result. Language-based methods analyse how speakers describe events. Contextual encoders can represent sentence meaning and word relationships. In multimodal settings, text can complement voice and facial information when one signal is

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

DOI: 10.17148/IARJSET.2026.131169

weak [3], [12]. The framework uses RoBERTa-based features and a Random Forest decision model. The selected encoder configuration is described in Section III. Linguistic deception detection approaches the problem through the content and organization of statements. Words, context, and semantic relationships can be represented by modern contextual encoders rather than simple frequency counts. The framework uses RoBERTa-base and selective adaptation to construct text features. This approach differs from an end-to-end fusion model because the final text decision is generated by a separate Random Forest classifier. Kamboj et al. explored multimodal political deception detection [12], demonstrating the relevance of language and nonverbal behaviour in another communication setting. Such work also illustrates the importance of task context: political statements, courtroom testimony, and controlled interviews can differ in vocabulary, speaking style, and recording structure. Reported results should therefore be compared with attention to the underlying datasets and evaluation procedures. Audio-based methods examine measurable properties of the voice rather than the words alone. Fundamental frequency, energy, spectral characteristics, and voice quality can describe how a statement is delivered. openSMILE toolkit supports repeatable extraction of such descriptors [17]. In the study, eGeMAPS and EmoBase provides complementary groups of acoustic measurements. These measurements can respond to emotion, speaking style, or microphone conditions as well as deception. Text analysis provides another view of the same statement. Unlike the audio signal, a transcript preserves lexical and contextual information but removes most prosodic detail. Contextual language models can represent relationships among words beyond isolated word counts. The framework uses RoBERTa representations with parameter-efficient adaptation and a Random Forest classifier. Its linguistic model is evaluated as an independent branch rather than as a jointly trained audio-text network. The distinction between spoken content and speaking behaviour matters for multimodal design. A transcript can remain informative when audio quality is poor, whereas speech features may add information not visible in the transcript. At the same time, transcription mistakes can affect linguistic representations. These differences support examining the modalities separately before combining their decisions [3], [12].

C. *Multimodal Fusion*

Multimodal systems can join features before classification or combine decisions after separate classifiers make predictions. Baltrušaitis et al. described these fusion families and their main design challenges [20]. Touma et al. investigated voting-based deception detection [2]. Sen et al. evaluated multimodal representations on trial data [3]. Kang et al. used global and local feature fusion with attention [7]. Late fusion offers a practical benefit for heterogeneous inputs. The video branch can use deep image embeddings, the audio branch can use engineered descriptors, and the text branch can use contextual representations. Their outputs become comparable after conversion to class probabilities. The present study uses accuracy-proportional soft voting rather than a new end-to-end neural fusion network. Research on the Bag-of-Lies dataset [5] and the DOLOS benchmark [4] also shows the need for careful evaluation across recording conditions. These datasets are discussed as background only. The reported experiments in this paper concern the Real-Life Trial dataset. Multimodal learning has been surveyed in terms of representation, alignment, fusion, and evaluation challenges [20]. In early fusion, features are integrated before a final classifier is trained. In late fusion, each modality yields a decision or probability that is combined at a later stage. The choice depends on the available data, the reliability of cross-modal alignment, and the complexity that the training set can support. Touma et al. investigated voting-based multimodal deception detection [2], while Sen et al. studied multimodal analysis on real-life trial data [3]. Kang et al. explored global and local feature fusion with multi-head attention [7]. These works show that multimodal integration can be implemented through relatively simple decision rules or through learned feature interactions. They also provide motivation for evaluating a transparent weighted average as a competitive, low-complexity baseline. Gupta et al. introduced the Bag-of-Lies multimodal dataset [5], and Guo et al. described the DOLOS audio-visual benchmark [4]. These resources broaden the available research settings beyond Real-Life Trial recordings. They are relevant for future validation, but the experiments presented here were conducted on the Real-Life Trial dataset; results on other datasets must not be implied. Multimodal deception detection can combine information at several stages. Early fusion joins modality features before a shared classifier, whereas late fusion combines predictions from separate classifiers. Baltrusaitis et al. describe broader multimodal challenges involving representation, alignment, and fusion [20]. The choice of fusion stage changes the amount of interaction learned between modalities and the difficulty of training the model. Voting-based work by Touma et al. illustrates decision-level combination [2]. Sen et al. investigate multimodal deception detection on real-life trial data [3], while Kang et al. study global and local feature fusion with multi-head attention [7]. These methods provide different ways to use complementary signals. Their reported scores should be compared cautiously because evaluation partitions and feature preparation may differ. The Bag-of-Lies dataset introduced by Gupta et al. [5] and the DOLOS benchmark described by Guo et al. [4] broaden the available research settings. They illustrate that data collection procedures and recording conditions vary across benchmarks. The current study does not report new experiments on those datasets. Its quantitative results come from the Real-Life Trial dataset. A gap remains between sophisticated multimodal architectures and simple, transparent combination rules. Complex feature fusion may learn cross-modal interactions, but it often requires more training data and careful tuning. Accuracy-weighted late fusion offers a compact alternative. It assigns fixed weights to independently trained classifiers and allows the contribution of each modality to be inspected. The present paper evaluates that design using the results documented in the experimental records.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

DOI: 10.17148/IARJSET.2026.131169

III. PROPOSED METHODOLOGY

The proposed framework has three parallel processing branches. Each branch converts one input modality into a fixed-length representation and estimates the probability of each class. A final decision module combines the probability estimates. The design follows the actual processing steps documented in the experimental records rather than treating all modalities as one shared neural network.

A. Problem Formulation

Let a recording be represented by X = (V, A, T), where V denotes video frames, A denotes the speech signal, and T denotes the transcript. Each modality-specific model outputs a probability for the truthful class. The dataset contains two classes: deceptive (0) and truthful (1), as defined in the audio-processing and fusion implementation. The final decision uses a fusion threshold.

The three feature families also differ in how they represent time. Acoustic functionals compress the full utterance into summary statistics, the visual branch averages frame-level embeddings, and the linguistic branch preserves token context during encoding before constructing a fixed-length representation. These choices reduce the complexity of the final classifiers, but they can suppress fine-grained temporal coordination between facial movement, speech delivery, and word choice. Feature standardization is especially relevant to the acoustic branch because the descriptors use different units and numerical ranges. A scaling transformation estimated from training data can help the MLP learn from multiple feature families without allowing high-magnitude descriptors to dominate merely because of scale. The source methodology describes training-based scaling; it should be implemented separately within each evaluation split when the experiment is reproduced.

B. Framework Overview

TABLE I PROCESSING STAGES OF THE THREE MODALITIES

Visual branch	Audio branch	Linguistic branch
Video → 16 sampled frames	Audio waveform	Transcript text
Haar face detection → crops	openSMILE eGeMAPSv02 + EmoBase	RoBERTa tokenizer → encoder
SE-ResNet50 → mean pooling	1076-dimensional acoustic vector	IA³ adaptation → embedding
Random Forest → P_v	MLP → P_a	Random Forest → P_t

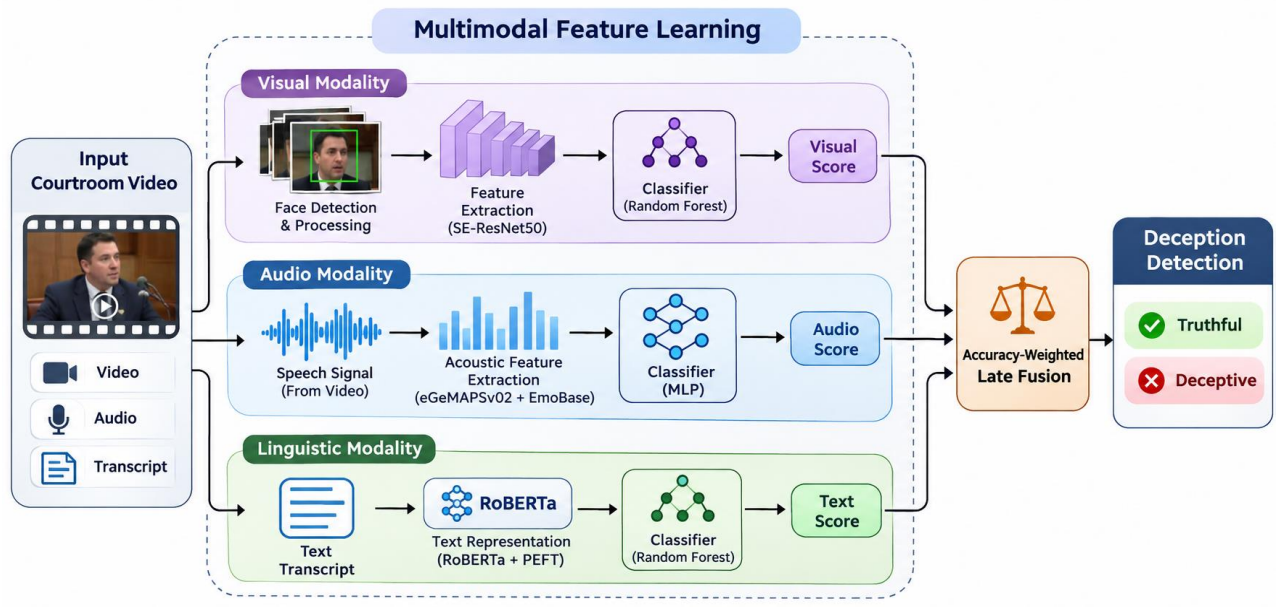


Fig. 1. Proposed three-branch processing and decision-level fusion workflow (branches are trained independently, not jointly).

C. Data Alignment

All three predictions must refer to the same original recording before fusion. The framework uses per-sample feature files and modality predictions. Recording identifiers provide the basis for alignment. A valid fusion experiment requires a one-to-one match among labels, sample identities, and probability columns. The exact alignment checks are not fully specified; no additional unverified alignment steps are assumed. The complete dataset contains 121 trial clips. The class

 This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

DOI: 10.17148/IARJSET.2026.131169

distribution is nearly balanced, with 60 truthful and 61 deceptive samples [1]. This small sample size makes leakage prevention and independent testing especially important. Feature normalization and classifier fitting should occur within training folds. A test fold should not influence fusion weights or threshold selection.

D. Visual Feature Extraction

Face Detection and Sampling The visual pipeline samples up to 16 frames at evenly spaced positions in each video. It applies the OpenCV Haar Cascade detector to each sampled frame. The detector uses a scale factor of 1.1, five minimum neighbours, and a minimum face size of 30 × 30 pixels. When a frame contains more than one detected face, the pipeline keeps the largest face region. It discards sampled frames without a detected face.

$$t_i = round\,[i(T-1)/(N-1)],\ i = 0, \ldots, N-1;\ N = 16$$

Here, T is the total number of frames and t_i is a sampled position. The detected region is cropped and resized for feature extraction. The resulting number of valid frames may be lower than 16. The visual preprocessing protocol specifies that videos without any detected face are excluded from the visual analysis. It does not provide a final exclusion count in the reported fusion table. Deep Facial Representation

The study compares a pretrained ResNet50 representation with an SE-ResNet50 variant. The models operate as frozen feature extractors. Each accepted frame produces a 2048-dimensional feature vector. ResNet50 uses residual learning to represent appearance [15]. The SE mechanism recalibrates channel responses [16]. The model uses the SE variant as emotion-oriented, but it does not establish that each selected feature corresponds to a specific facial emotion. The preprocessing includes a 224 × 224 centre crop and ImageNet normalization. The preprocessing also includes resizing to 260 pixels before centre cropping for the SE variant. It does not fine-tune these visual backbones on the trial dataset. Temporal Pooling and Classification A recording may contain several accepted frames. The pipeline averages their embeddings to create one vector for the entire clip. This reduces the number of observations passed to the classifier. It also removes explicit frame order, so the visual model does not directly learn temporal motion.

$$\bar{x}_v = (1/k)\,\Sigma_{j=1}^{k} f(frame_j),\ \bar{x}_v \in \mathbb{R}^{2048}$$

The experiments evaluate Random Forest and Logistic Regression configurations, together with a ResNet-based comparison. The strongest visual configuration use SE-ResNet emotion-oriented features with Random Forest and reports 0.76 accuracy. A 5-fold stratified protocol is specified for the visual branch. The study also reports 0.839 AUC for this configuration. The visual results suggest that channel-reweighted image embeddings provide useful information. However, the modest sample size and absence of subject-independent validation in the study prevent a claim that the method is identity-independent.

E. Acoustic and Linguistic Processing

Acoustic Descriptors The audio pipeline extracts two complementary openSMILE feature sets [17]. eGeMAPSv02 provides 88 functionals that summarize properties such as fundamental frequency, energy, spectral shape, and voice quality. EmoBase provides 988 functionals, including cepstral and other low-level descriptor statistics. The two vectors are concatenated into a 1076-dimensional representation for each recording.

$$x_a = [x_{eGeMAPS} \,//\, x_{emoBase}],\ dim(x_a) = 88 + 988 = 1076$$

The experiments evaluate several conventional classifiers for the acoustic representation. The strongest model is a multilayer perceptron with one hidden layer of 100 ReLU neurons. It uses Adam, up to 500 iterations, and random seed 42. A Standard Scaler standardizes the features. The methodology states that scaling is fitted on the training portion and then applied to the held-out portion. Linguistic Encoder

The text branch reads transcripts and tokenizes them with a maximum length of 256 tokens. The backbone is RoBERTa-base with 12 Transformer layers and a hidden dimension of 768. The lower eight layers remain frozen. The upper four layers, the pooler, and IA^3 adapter components receive task-specific updates. The model uses ramp-weighted hidden-state aggregation before classification. The reported text training settings include 25 epochs, batch size 8, AdamW, and cross-entropy loss. Different parameter groups receive different learning rates. The final reported classifier is Random Forest. These settings should not be confused with the audio MLP configuration. Text Model Reporting Limitation The intermediate feature dimension after the 768-dimensional RoBERTa representation is not fully specified. The text branch uses RoBERTa-based feature extraction with selective adaptation, followed by Random Forest classification. No additional projection or feature-selection operation is assumed.

Class-label consistency between the text and audio pipelines requires verification. In this paper, the displayed confusion-related class terminology follows the audio and fusion convention: deceptive = 0 and truthful = 1. The original code or prediction files should be checked before publication to confirm consistent label mapping across modalities.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

DOI: 10.17148/IARJSET.2026.131169

F. Accuracy-Weighted Late Fusion

Probability Combination Each trained model returns a probability for the truthful class. Let P_t, P_v, and P_a be the text, visual, and audio probabilities. The fusion layer calculates their weighted sum. The weights are nonnegative and sum to one. Thus, the result remains within the probability interval.

$$P_{fusion} = w_t P_t + w_v P_v + w_a P_a$$

$$w_i = Acc_i / (Acc_t + Acc_v + Acc_a),\ \Sigma_i w_i = 1$$

The experiments report weights of 0.3116 for text, 0.3333 for video, and 0.3551 for audio. The audio branch receives the largest weight because it obtains the highest individual accuracy. The text branch receives the smallest weight, but it can still change the final classification when its probability differs from the other branches. Weighted averaging can be contrasted with simple probability averaging. Under equal averaging, all branches contribute one third of the final score regardless of their standalone performance. Under accuracy-proportional weighting, a branch with higher reported accuracy receives a larger coefficient. The relative weights in this study are close enough that all three branches still influence borderline predictions. Consequently, the method is not equivalent to selecting only the strongest audio classifier.

TABLE II FUSION WEIGHTS REPORTED FOR EACH MODALITY

Modality	Fusion weight	Role in fusion
Text	0.3116	Linguistic probability
Visual	0.3333	Facial representation probability
Audio	0.3551	Speech probability

Threshold Selection

The reported weighted method applies a decision threshold of 0.50. A fused score at or above the threshold is assigned to the truthful class, and a lower score is assigned to the deceptive class. Different decision thresholds are specified for the alternative fusion models. It does not fully document whether the fusion threshold and modality weights were selected independently within training folds. This uncertainty matters because using held-out labels to tune a threshold can overestimate performance. The fusion threshold is distinct from the modality weights. Weights control how the three probabilities are combined, while the threshold converts the combined score into a binary decision. Changing either component can alter the confusion matrix. Because the experimental records report a selected threshold of 0.50, the manuscript preserves that value rather than assuming the conventional 0.50 cutoff. For an unbiased future experiment, both the weighting procedure and threshold selection should be confined to training or validation data. Illustrative Decision Consider a recording for which the text and visual branches favour the truthful class while audio favours deception. The weighted average may remain above the threshold because the two supportive probabilities together outweigh the audio probability. An illustrative example demonstrates a worked example of this situation. This illustrates the mechanism, not a universal guarantee that fusion corrects an incorrect model. The method has low computational cost after the three classifiers produce their probabilities. It requires only three multiplications, two additions, and a threshold comparison per recording. The main computational burden remains in feature extraction and classifier inference. The method is easy to inspect because each modality has a visible weight and output probability.

IV. EXPERIMENTAL SETUP AND RESULTS

A. Dataset and Measures

The Real-Life Trial dataset contains 121 courtroom video clips, including 61 deceptive and 60 truthful recordings [1]. The experiments assess separate models for visual, acoustic, and linguistic models, followed by fusion of the three outputs. It uses stratified cross-validation in the visual and audio pipelines. The available description does not establish a single matched outer-fold protocol for all branches and the final fusion. Accuracy measures the fraction of correct predictions. Precision measures the fraction of positive predictions that are correct. Recall measures the fraction of actual positive cases detected. F1 combines precision and recall. AUC summarizes ranking quality across decision thresholds. The results tables report these measures for individual classifiers and fusion methods.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

$$Precision = TP/(TP + FP);\ Recall = TP/(TP + FN)$$

$$F1 = 2 \times Precision \times Recall/(Precision + Recall)$$

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

DOI: 10.17148/IARJSET.2026.131169

Reported Implementation Choices

TABLE III REPORTED IMPLEMENTATION CHOICES

Branch	Features	Selected classifier	Main settings
Visual	2048-D mean-pooled SE-ResNet	Random Forest	16 sampled frames; 5-fold CV
Audio	1076-D openSMILE	MLP	100 hidden units; Adam; seed 42
Text	RoBERTa + IA^3	Random Forest	256 tokens; top layers adapted
Fusion	Three class probabilities	Weighted average	0.3116 / 0.3333 / 0.3551

Evaluation Cautions A rigorous evaluation should compute the weights from training-fold performance only. The threshold should also be chosen without seeing the outer test fold. Any scaler, classifier, or parameter-selection step should be fitted inside the relevant training partition. These are evaluation requirements rather than new experiments performed for this paper. The available evaluation records include overall performance and rescue analysis, but do not establish all fold assignments, subject identities, or complete out-of-fold probabilities. Therefore, the results are reported without claiming an additional independent replication or statistical significance test.

B. Results and Comparative Analysis

Single-Modality Results

TABLE IV SINGLE-MODALITY RESULTS

Model	Accuracy	AUC	F1
Audio: MLP	0.81	0.8667	0.8122
Visual: SE-ResNet + RF	0.76	0.839	0.766
Text: RoBERTa + RF	0.71	0.770	0.696

The audio MLP performs best among the individual modalities. The visual branch ranks second, followed by the linguistic branch. The text result shows that semantic representations alone do not solve this task on the selected dataset. These values summarize the individual classifier evaluations. The visual result differs slightly from the 76.7% value used for weight computation in the fusion-weight calculation; both values are preserved in their original contexts. Fusion Results

TABLE V FUSION RESULTS

Fusion method	Threshold	Accuracy	AUC	Macro F1
Accuracy-weighted average	0.50	0.8678	0.9008	0.8678
Simple average	0.51	0.8595	0.8992	0.8595
Sigmoid neuron grid 6D	0.46	0.8595	0.9074	0.8595

The weighted average obtains the highest reported accuracy and F1 among the three fusion methods. The sigmoid neuron grid produces the highest AUC, but not the highest accuracy. It would therefore be inaccurate to attribute the 0.9074 AUC to the proposed weighted method. The weighted method has 0.9008 AUC and 86.78% accuracy. Improvement Over the Audio Baseline The experiments report 98 correct predictions for audio and 105 correct predictions for weighted fusion out of 121 recordings. This represents seven additional correct predictions and an accuracy gain of 5.79 percentage points. The improvement is descriptive. Without a documented paired statistical test, it should not be described as statistically significant.

TABLE VI AUDIO BASELINE VERSUS WEIGHTED FUSION

Configuration	Correct / total	Accuracy
Audio MLP	98 / 121	80.99%
Weighted fusion	105 / 121	86.78%

The difference between 81.03% in the classifier summary and 80.99% in the pooled prediction comparison reflects different reported evaluation summaries. Both values are identified according to their evaluation context.

V. ERROR ANALYSIS AND DISCUSSION

Rescue Analysis The analysis compares audio-only decisions with the weighted fusion decisions on the same 121 recordings. Both methods classify 94 recordings correctly. Fusion changes 11 audio errors into correct decisions. It also

 This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

DOI: 10.17148/IARJSET.2026.131169

changes four correct audio predictions into errors. Twelve recordings remain incorrect under both methods. The net gain is seven correct predictions.

TABLE VII RESCUE ANALYSIS OF AUDIO AND FUSION PREDICTIONS

Prediction relationship	Recordings
Both audio and fusion correct	94
Audio wrong, fusion correct	11
Both audio and fusion wrong	12
Audio correct, fusion wrong	4

Among the 11 corrected recordings, six belong to the deceptive class and five to the truthful class. The corrections therefore appear on both sides of the class distribution. However, fusion also introduces new errors. The results show a trade-off rather than error-free improvement. Comparison with Studies Listed in the study

TABLE VIII COMPARISON WITH EXISTING METHODS

Existing methods	Reported accuracy
Yildirim et al. [9]	74.17%
Mathur and Matarić [23]	75.00%
Ngo et al. [22]	68.00%
Sen et al. [3]	83.05%
Proposed weighted fusion	86.78%

The comparison presents reported values from the cited publications. It is contextual rather than a controlled head-to-head experiment. Differences in preprocessing, partitions, subject overlap, and evaluation protocols can change results. The current study therefore cannot establish superiority over every previous method under identical conditions. Interpretation The findings suggest that different modalities make partly different errors. Audio carries useful prosodic information, while facial representations and transcript features can shift the final probability when audio is misleading. Weighted averaging also reduces dependence on a single classifier. At the same time, fixed global weights may not reflect the reliability of a modality for each individual recording. The strongest practical contribution is the straightforward integration of heterogeneous representations. The main limitations are the small benchmark, uncertain common evaluation protocol, and lack of a verified subject-independent experiment. These issues should be addressed before describing the framework as ready for high-stakes deployment.

VI. LIMITATIONS AND FUTURE WORK

The study has several limitations. First, the Real-Life Trial benchmark contains only 121 recordings. This is useful for controlled comparison but insufficient to establish performance across languages, speakers, courts, and recording conditions. Second, the visual representation averages selected frame embeddings and does not preserve explicit motion. Third, the fusion weights are global constants. They cannot adjust to a missing face, poor audio, or an unreliable transcript on a specific recording.

A further limitation concerns the evaluation description. The experiments report stratified validation for individual branches, but the final probability-fusion protocol is not fully documented. A future experiment should save out-of-fold predictions from every branch, choose weights and thresholds inside training partitions, and report an untouched outer-fold evaluation. Subject-grouped validation would help assess whether the model generalizes to unseen speakers. Future research may compare the current weighted rule with equal averaging, calibrated stacking, and uncertainty-aware fusion under the same splits. It may also examine whether learned temporal visual features improve over mean-pooled image embeddings. Larger datasets such as Bag-of-Lies [5] and DOLOS [4] provide possible settings for external evaluation, but no such experiment is claimed here. Deception detection can affect reputations and decisions about individuals. A classifier score should not be treated as proof that a person is lying. Bias, dataset context, privacy, and error costs require careful review. The system is best viewed as an experimental research tool, not a substitute for independent evidence or human judgment.

VII. CONCLUSION

This paper presented an accuracy-weighted late-fusion framework for deception detection using visual, acoustic, and linguistic information. The visual branch uses SE-ResNet50 representations with Random Forest. The audio branch combines openSMILE eGeMAPSv02 and EmoBase features with an MLP. The text branch uses RoBERTa-based features and Random Forest. A weighted average combines the three class probabilities. On the Real-Life Trial dataset, the study reports 86.78% accuracy, 0.9008 AUC, and 0.8678 macro F1 for the weighted fusion method. Fusion improves

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

DOI: 10.17148/IARJSET.2026.131169

accuracy by 5.79 percentage points over the strongest audio-only baseline and corrects 11 of its errors while introducing four new ones. These results indicate the potential value of complementary evidence. More rigorous matched-fold and subject-independent evaluation is needed to confirm the reliability and generalization of the method.

REFERENCES

[1]. V. Pérez-Rosas, M. Abouelenien, R. Mihalcea and M. Burzo, "Deception detection using real-life trial data", Proc. ACM Int. Conf. Multimodal Interaction (ICMI), pp. 59–66, 2015.

[2]. L. Touma, M. Al Horani, M. Tailouni, A. Dahabiah and K. Al Jallad, "Voting-based multimodal automatic deception detection", arXiv preprint, arXiv:2307.07516, 2023.

[3]. M. U. Şen, V. Pérez-Rosas, B. Yanıkoğlu, M. Abouelenien, M. Burzo and R. Mihalcea, "Multimodal deception detection using real-life trial data", IEEE Trans. Affect. Comput., vol. 13, no. 1, pp. 306–319, 2022.

[4]. X. Guo, N. M. Selvaraj, Z. Yu, A. W.-K. Kong, B. Shen, and A. Kot, "Audio-visual deception detection: DOLOS dataset and parameter-efficient crossmodal learning," *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, pp. 22078–22088, 2023.

[5]. V. Gupta, M. Agarwal, M. Arora, T. Chakraborty, R. Singh and M. Vatsa, "Bag-of-Lies: A multimodal dataset for deception detection", Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 83–90, 2019.

[6]. J.-T. Yang, G.-M. Liu and S. C.-H. Huang, "Constructing robust emotional state-based feature with a novel voting scheme for multi-modal deception detection in videos", arXiv preprint, arXiv:2104.08373, 2021.

[7]. J. Kang, W. Qu, S. Cui and X. Feng, "Deception detection algorithm based on global and local feature fusion with multi-head attention", Proc. 3rd Int. Conf. Image Processing and Media Computing (ICIPMC), pp. 162–168, 2024.

[8]. P. K. Sehrawat, R. Kumar, N. Kumar and D. K. Vishwakarma, "Deception detection using a multimodal stacked Bi-LSTM model", Proc. Int. Conf. Innovative Data Communication Technologies and Application (ICIDCA), pp. 318–326, 2023.

[9]. S. Yildirim, M. S. Chimeumanu and Z. A. Rana, "The influence of micro-expressions on deception detection", Multimedia Tools Appl., vol. 82, no. 19, pp. 29115–29133, 2023.

[10]. A. Vrij, Detecting Lies and Deceit: Pitfalls and Opportunities, 2nd ed., John Wiley & Sons, 2008.

[11]. P. Ekman, Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage, W. W. Norton & Company, 2009.

[12]. M. Kamboj, C. Hessler, P. Asnani, K. Riani and M. Abouelenien, "Multimodal political deception detection", IEEE MultiMedia, vol. 28, no. 1, pp. 94–102, 2021.

[13]. P. Ekman and W. V. Friesen, Facial Action Coding System, Consulting Psychologists Press, 1978.

[14]. A. Derakhshan, M. Mikaeili, T. Gedeon and A. M. Nasrabadi, "Identifying the optimal features in multimodal deception detection", Multimodal Technol. Interact., vol. 4, no. 2, Art. no. 25, 2020.

[15]. K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition", Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

[16]. J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks", Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141, 2018.

[17]. F. Eyben, M. Wöllmer and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor", Proc. ACM Int. Conf. Multimedia, pp. 1459–1462, 2010.

[18]. J. Zhang, S. I. Levitan and J. Hirschberg, "Multimodal deception detection using automatically extracted acoustic, visual, and lexical features", Proc. Interspeech 2020, pp. 359–363, 2020.

[19]. L. Breiman, "Random forests", Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[20]. T. Baltrušaitis, C. Ahuja and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy", IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, 2019.

[21]. C. F. Bond and B. M. DePaulo, "Accuracy of deception judgments", Pers. Soc. Psychol. Rev., vol. 10, no. 3, pp. 214–234, 2006.

[22]. L.-M. Ngô, W. Wang, B. Mandira, S. Karaoğlu, H. Bouma, H. Dibeklioğlu and T. Gevers, "Identity unbiased deception detection by 2D-to-3D face reconstruction", Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV), pp. 145–154, 2021.

[23]. L. Mathur and M. J. Matarić, "Affect-aware deep belief network representations for multimodal unsupervised deception detection", Proc. 16th IEEE Int. Conf. Automatic Face and Gesture Recognition (FG), pp. 1–8, 2021.

[24]. C. Zhou, Y. Zhang, C. Lin and S. Zhou, "A deception detection model by using integrated LLM with emotion features", Scientific Reports, vol. 15, Art. no. 32135, 2025.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

DOI: 10.17148/IARJSET.2026.131169

BIOGRAPHY



Maheta Hardikkumar Harishbhai: Maheta Hardikkumar Harishbhai is a Ph.D.research scholar from Gujarat Technological University, Ahmedabad, Gujarat, India. He received his B.Tech. degree in Information Technology from Nirma University, Ahmedabad, Gujarat, India, in 2009, and his M.Tech. degree in Information Technology from Dharmsinh Desai University, Nadiad, Gujarat, India, in 2014. He is currently serving as an Assistant Professor in the Department of Information Technology at Shantilal Shah Engineering College, Bhavnagar, Gujarat, India. He has over 16 years of teaching experience. His research interests include artificial intelligence, machine learning, deep learning, computer vision, image processing, and data science. He has published several research papers in reputed national and international conferences and journals in the fields of artificial intelligence and machine learning.



Dr. Dineshkumar Bhagwandas Vaghela*: Dr. Dineshkumar Bhagwandas Vaghela received his Ph.D. in Computer Engineering from Gujarat Technological University, Chandkheda, Gujarat, India, in 2017. He completed his M.E. in Computer Engineering from Sardar Patel University, Vallabh Vidyanagar, Gujarat, India, in 2010, and his B.E. in Information Technology from Gujarat University, Ahmedabad, Gujarat, India, in 2004. He is currently serving as an Assistant Professor in the Department of Information Technology at Shantilal Shah Engineering College, Bhavnagar, Gujarat, India. He has 21 years of teaching experience. His research interests include computer vision, image processing, machine learning, cloud computing, big data analytics, and the Internet of Things (IoT). He has published several research papers in reputed national and international conferences and journals in the areas of computer vision, image processing, machine learning, and cloud computing. **He is the corresponding author of the present research paper.**

.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License