# Multimodal Fusion for Deception Detection

Ivan Galvan Gomez
*Universidad Panamericana*
Aguascalientes, Mexico
0246325@up.edu.mx

***Abstract*—The effectiveness of multimodal integration for lie detection was examined, addressing a critical need in legal and forensic contexts. The study used the ATSFace dataset, which provides precomputed features: 128-dimensional FaceNet vectors per video frame, 128-dimensional BERT sentence embeddings per clip (obtained by averaging all sentence-level embeddings), and averaged MFCC_0.2 audio features. A modular approach was employed, where separate models were trained independently for the visual and auditory modalities. The textual modality was incorporated through preprocessed averaged embeddings without further fine-tuning. Optimized representations from each channel were subsequently combined using a gated fusion mechanism to train a final classifier. This strategy enabled improved integration of heterogeneous signals and facilitated a detailed evaluation of individual modality contributions, offering evidence of increased robustness and interpretability compared to conventional fusion techniques.**

***Index Terms*—Lie Detection, Multimodal Integration, ASTFace, FaceNet, BERT, Feature Fusion**

## I. Introduction

The ability to detect deception is critical in a variety of high-stakes environments, particularly in legal, forensic, and security contexts. False statements can have severe consequences, and accurately identifying deception can support investigations, improve the objectivity of the courtroom, and prevent miscarriages of justice. Traditional lie detection techniques, such as polygraph tests, focus on physiological signals such as heart rate or skin conductivity. However, these unimodal approaches suffer from low reliability and can be manipulated by subjects [1].

In response to these limitations, researchers have explored computational approaches using deep learning. Recent work has shown that individual modalities—such as facial expressions, speech, or linguistic content—can reveal subtle cues of deception [2], [3]. Nevertheless, these unimodal systems are often unable to fully capture the complexity of human communication. The inherent ambiguity and subjectivity of single-channel signals make them vulnerable to noise, overfitting, and contextual misinterpretation.

To address these weaknesses, multimodal approaches have emerged. By integrating visual, textual, and auditory signals, multimodal models aim to improve robustness and accuracy. While early, late, and hybrid fusion techniques have been proposed [4], they often involve tightly coupled architectures that hinder interpretability and do not allow individual analysis of each modality's contribution [5]. Additionally, existing datasets often lack the diversity and structure needed for a comprehensive multimodal analysis [6].

This research proposes a modular and interpretable multimodal approach to lie detection, leveraging the ATSFace dataset, which provides aligned audio, visual, and transcription features. In contrast to traditional fusion strategies, the proposed method trains specialized models independently for each modality, extracts optimized features, and performs horizontal fusion for final classification. This separation enables each modality to be processed according to its unique characteristics and facilitates independent or joint analysis of their contributions.

The key contributions of this work are as follows:

- A modular pipeline is proposed in which visual, audio, and textual features are independently processed using deep neural networks and subsequently fused into a compact representation for deception classification.
- The effectiveness of the approach is demonstrated on the ATSFace dataset, with preliminary results showing promising accuracy using only visual and textual modalities.
- A foundation is provided for further interpretability and per-modality analysis, enabling the identification of which types of cues—visual, auditory, or verbal—are most predictive of deceptive behavior.

The code and resources for this project are available at: https://github.com/Ivan10121/Multimodal-Fusion.git

## II. Previous Work

Traditional lie detection methods have long relied on physiological measurements, such as polygraph tests that monitor heart rate, blood pressure, and skin conductivity [1], [7]. Although widely used, these approaches often suffer from limited accuracy and can be affected by countermeasures.

### *A. Deep Learning Approaches*

Recent advances in deep learning have introduced alternative techniques for lie detection:

- **Visual Analysis:** Several studies have demonstrated that Convolutional Neural Networks (CNNs) and pre-trained models like FaceNet [8] can effectively extract features from facial images, capturing microexpressions and other non-verbal cues associated with deceptive behavior [2].
- **Natural Language Processing:** The development of language models such as BERT [3] has enabled researchers to analyze semantic and contextual nuances in textual data, contributing to a better understanding of verbal indicators of deception [9].

### *B. Multimodal Fusion Techniques*

To overcome the limitations of unimodal systems, multimodal approaches that integrate audio, visual, and textual cues have been proposed:

- **Data Integration:** Various fusion strategies—including early, late, and hybrid methods—have been explored to combine heterogeneous data sources, aiming to achieve a more comprehensive analysis of deceptive behavior [4].
- **Existing Challenges:** Despite progress, current multimodal fusion techniques often do not fully address the specific contributions of each modality or adapt easily to the diverse nature of the data [5].

### *C. Research Gaps and Contributions*

While the literature offers promising methodologies, some gaps remain:

- **Dataset Diversity:** Many available datasets do not provide sufficient multimodal variety, limiting the scope of analysis [6].
- **Modular Fusion Strategies:** Few studies have proposed modular frameworks that allow independent training of each modality, which can help in understanding and optimizing their individual contributions [10].

In this work, these gaps are addressed by leveraging the unique characteristics of the ATSFace dataset. The proposed approach:

- Trains specialized models independently for the audio, video, and transcription modalities, enabling tailored feature extraction for each type of data.
- Employs a fusion strategy to integrate the optimized features, allowing for a clearer analysis of the contribution of each modality.

## III. Methodology

### *A. Dataset*

The ATSFace dataset authors did not release the raw video files. Instead, the published dataset consists solely of the precomputed feature vectors.

The dataset creation protocol consisted of three phases::

1) **Initial Questions:** Participants answered general questions about their school life and financial background.
2) **Fictitious Narratives:** Participants were asked to fabricate stories on selected topics, such as their major, club experiences, internships, travel anecdotes, and personal hobbies.
3) **Truthful Narratives:** Participants then provided honest accounts about a different chosen topic.

*Recording Setup:* All sessions were recorded using an iPhone 14 Pro at 1080p resolution and 30 fps. Subjects responded to a moderator's prompts in Chinese, exhibiting a variety of facial expressions and speech behaviors.

*Composition:* The dataset contains:

- **Total Videos:** 309 clips (147 deceptive, 162 truthful).
- **Duration:** Average of 23.32 s (range: 10.53 s to 49.73 s).
- **Speakers:** 36 unique individuals (23 male, 13 female).
- **Transcripts:**
  - Total words: 35,069 (1,403 unique).
  - Average words per transcript: 113.
  - Generated via CapCut ASR, retaining fillers and repetitions.

*Feature Extraction:* To facilitate multimodal analysis, the following feature sets were precomputed:

- **Visual:** Face embeddings extracted using RetinaFace for detection and a pretrained FaceNet model to produce 128-dimensional vectors (`facenet_128`).
- **Audio:**
  - `mfcc`: Frame-level Mel-frequency cepstral coefficients.
  - `mfcc_0.2`: Averaged MFCCs over non-overlapping 0.2 s windows.
- **Textual:** Transcriptions stored as SRT files, then tokenized using CKIP Lab's Chinese BERT:
  - chinese_bert_perword: Word-level embeddings (768 d).
  - chinese_bert_persentence: Sentence-level embeddings (768 d).

*SRT Format:* Each subtitle entry in the SRT files includes:

- A sequence number.
- Time codes in `hh:mm:ss,ms --> hh:mm:ss,ms` format.
- One or more lines of text.

This dataset provides a rich multimodal resource for deception detection research, balancing both deceptive and truthful speech across diverse topics.

### *B. Data Splitting*

An independent test set comprising 15% of the ATSFace data was first held out. The remaining 85% of the samples were then used for five-fold stratified cross-validation: within each fold, 80% of that subset (i.e. 68% of the total) was used for training and 20% (i.e. 17% of the total) for validation. The same fold indices were reused across visual, audio, and textual modality experiments, and likewise in the multimodal fusion stage, to ensure a fair and consistent comparison.

Feature extraction was performed on a per-fold basis. Specifically, for each cross-validation fold, the model achieving the highest validation accuracy was selected to extract feature representations from that fold's validation samples. These extracted features were saved and subsequently employed in the multimodal fusion stage. By strictly confining feature extraction to each fold's validation set—and by reusing consistent fold indices across modalities—this procedure prevented any leakage of information from held-out data into training or fusion, ensuring an unbiased evaluation.

### C. Visual Modality

Visual features were derived from the ASTFace dataset using the provided *facenet_128* embeddings. Each video clip is represented as a variable-length sequence of 128-dimensional vectors, one per frame. Ground truth labels were used to assign a binary classification target (deceptive or truthful) to each sequence.

To enable batch processing, sequences were zero-padded to the maximum length within each batch, and binary attention masks were constructed to distinguish valid tokens from padding.

A Transformer-based architecture was used for classification. It consists of:

- A learnable [CLS] token and positional embeddings of size 128,
- A Transformer encoder with 1 layers, 4 attention heads, a feed-forward size of 256, GELU activation, and dropout rate of 0.5,
- A classification head composed of two fully connected layers (128→64→2) with ReLU activations and dropout of 0.5,

Training and feature extraction followed the protocol detailed in Section III-B, using AdamW, an annealing scheduler, and early stopping on validation accuracy. Evaluation metrics included accuracy, precision, recall, and F1-score. The feature vectors extracted were saved in NumPy format for subsequent use in the multimodal fusion stage.

### D. Transcription Modality

Transcription features were obtained from precomputed sentence-level BERT embeddings provided in the ASTFace dataset. Each transcript consists of a variable-length sequence of 128-dimensional vectors, one per sentence. Ground truth labels were assigned in accordance with the deceptive or truthful nature of each clip. To produce fixed-length representations, mean pooling was applied across the sentence dimension, resulting in a single 128-dimensional vector per clip. This operation captures the average semantic content of the transcript and eliminates the need for padding or attention masking. These vectors were saved in NumPy format for use in the multimodal fusion pipeline.

### E. Audio Modality

Audio features were derived from frame-level embeddings provided in the ASTFace dataset, where each clip is represented as a variable-length sequence of 20-dimensional vectors. Ground truth labels were assigned based on whether the clip was classified as deceptive or truthful.

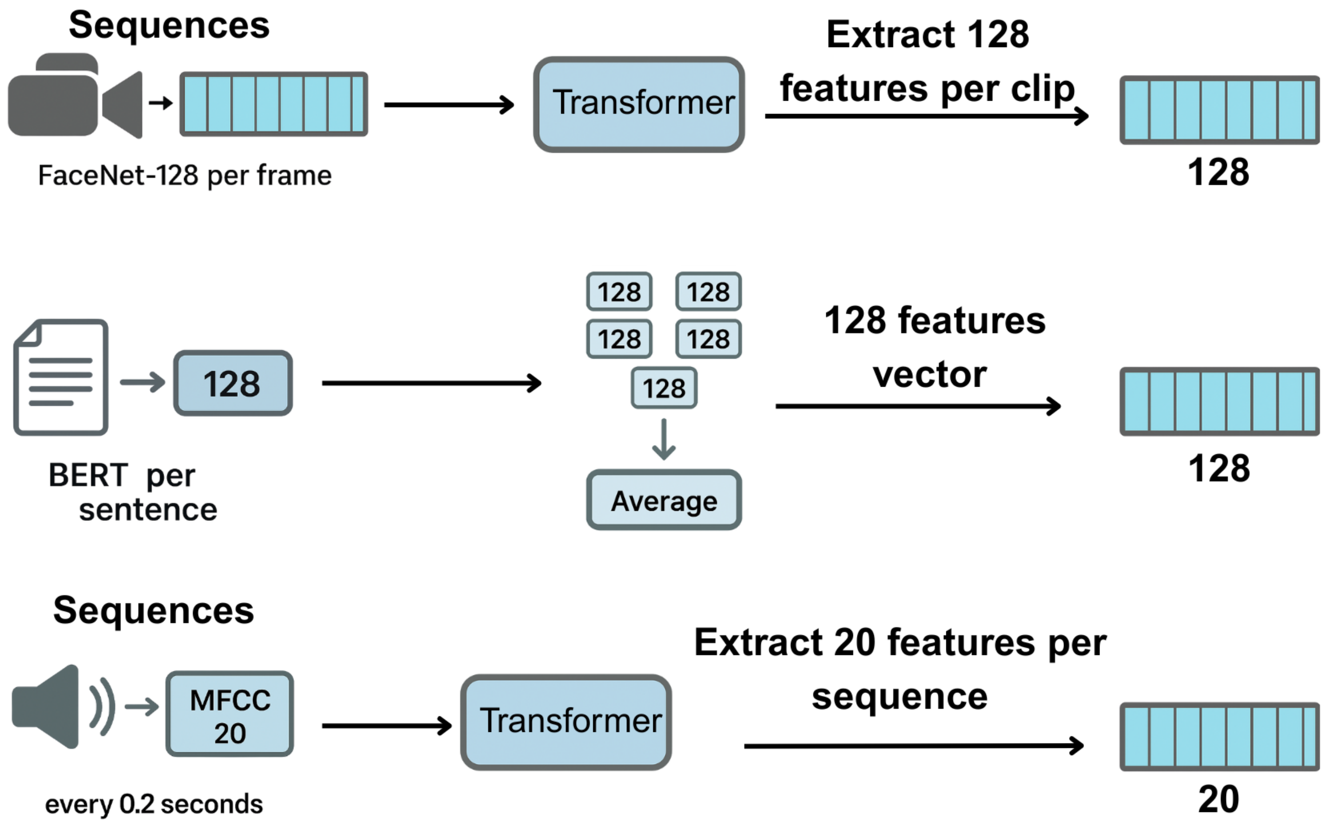


Fig. 1: Feature extraction pipeline for each modality. Each modality outputs a fixed-length embedding used for subsequent multimodal fusion.

To support batch processing, sequences were zero-padded to the maximum length within each batch, and binary attention masks were generated to distinguish real frames from padding.

A Transformer-based architecture was employed for classification. It consists of:

- A learnable [CLS] token and positional embeddings of size 20,
- A Transformer encoder with 1 layer, 4 attention heads, feedforward dimension 25, GELU activation, and dropout rate of 0.5,
- A classification head composed of two fully connected layers (20→64→2) with ReLU activations and dropout of 0.5,
- An auxiliary method to extract the 20-dimensional [CLS] embedding prior to classification.

Training and feature extraction followed the protocol detailed in Section III-B, using AdamW, an annealing scheduler, and early stopping on validation accuracy. Evaluation metrics included accuracy, precision, recall, and F1-score.

The feature vectors extracted were saved in NumPy format for subsequent use in the multimodal fusion stage.

### *F. Multimodal Fusion and Final Classifier*

Multimodal deception detection was performed by integrating the modality-specific features extracted from the visual, transcription, and audio pipelines. Each sample was represented by a 128-dimensional visual vector, a 128-dimensional transcription vector, and a 20-dimensional audio vector, along with a binary label indicating deception.

To enable joint processing, each modality is encoded into a 128-dimensional vector via a dedicated MLP. A softmax-gated fusion mechanism then combines these encoded representations into a single feature, which is fed to the final classifier.

The complete architecture consists of the following components:

- **Modality-specific projection layers:**
  - *Text:* $\text{Linear}(128 \to 128) \to \text{LayerNorm} \to \text{ReLU} \to \text{Dropout}$
  - *Video:* $\text{Linear}(128 \to 128) \to \text{LayerNorm} \to \text{ReLU} \to \text{Dropout}$
  - *Audio:* $\text{Linear}(20 \to 128) \to \text{LayerNorm} \to \text{ReLU} \to \text{Dropout}$

  These layers produce modality-aligned embeddings $t$, $v$, and $a$ of shape $(B, 128)$.
- **Softmax-gated fusion:**
  - The three modality embeddings are concatenated into a single vector of shape $(B, 384)$,
  - This concatenated vector is passed through a two-layer MLP:
    - $*$ $\text{Linear}(384 \to 128) \to \text{ReLU} \to \text{Dropout}$,
    - $*$ $\text{Linear}(128 \to 3)$,
  - A softmax is applied over the 3 logits to obtain weights $(w_t, w_v, w_a)$,
  - The fused representation is computed as a convex combination:

$$\text{fused} = w_t t + w_v v + w_a a \in \mathbb{R}^{128}.$$

- **Final classification head:**
  - $\text{Linear}(128 \to 64) \to \text{ReLU} \to \text{Dropout}$,
  - $\text{Linear}(64 \to 1)$, producing the final logit for binary classification.

Five-fold stratified cross-validation was conducted to preserve class distribution, using 80% of the data for training and 20% for testing in each fold. The model was trained using the AdamW optimizer, the loss function used was `BCEWithLogitsLoss`. A learning rate scheduler based on `ReduceLROnPlateau` was employed to reduce the learning rate. Performance was evaluated using accuracy, precision, recall, and F1-score.

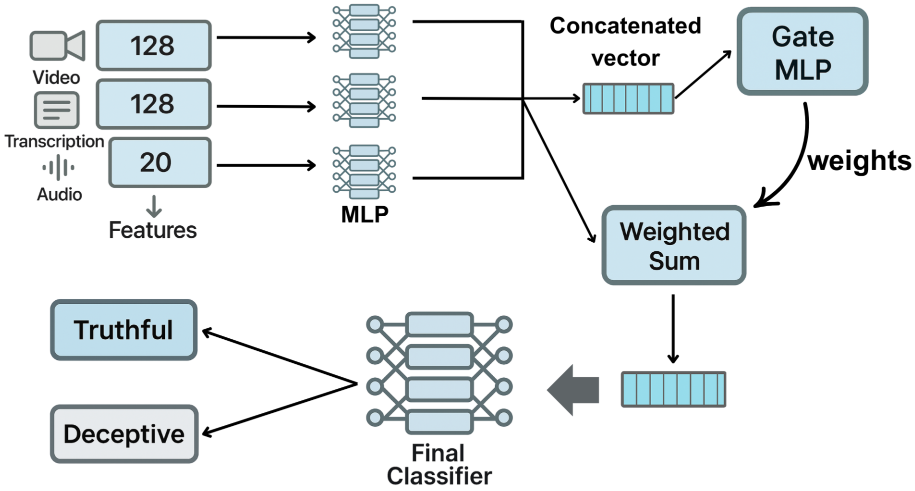


Fig. 2: General pipeline of gated multimodal fusion: features are extracted in parallel, projected into a shared latent space, and combined via a gating mechanism that adaptively weights each modality.

## IV. Training Configuration

All models were trained under a unified protocol to ensure comparability. Five-fold stratified cross-validation was applied for each modality, and the same fold indices were reused across visual and audio experiments to maintain consistency. Table I summarizes the main hyperparameters. Early stopping on the validation accuracy was applied with modality-specific patience values.

| Hyperparameter | Visual | Audio | Fusion |
|---|---|---|---|
| Optimizer | AdamW | AdamW | AdamW |
| Initial learning rate | 5e-4 | 5e-4 | 1e-3 |
| Weight decay | 1e-4 | 1e-4 | 0 |
| Scheduler | Cosine annealing | Cosine annealing | - |
| Batch size | 8 | 8 | 8 |
| Maximum epochs | 100 | 200 | 200 |
| Early-stopping patience | 30 | 30 | 10 |

TABLE I: Training hyperparameters for visual, audio, and fusion models.

## V. Results

### A. Visual Modality

In the visual modality (Table II), the model's performance exhibited considerable variability across folds, achieving an average accuracy of 46%, precision of 48%, recall of 60%, and an F1-score of 52%. These metrics indicate that the visual features alone provided limited reliability, suggesting that visual cues, in isolation, may not consistently capture deceptive behaviors effectively.

TABLE II: Performance on Visual Modality

| Metric | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Fold 1 | 0.51 | 0.52 | 0.8 | 0.63 |
| Fold 2 | 0.42 | 0.45 | 0.36 | 0.4 |
| Fold 3 | 0.61 | 0.58 | 0.92 | 0.71 |
| Fold 4 | 0.44 | 0.48 | 0.56 | 0.51 |
| Fold 5 | 0.36 | 0.39 | 0.36 | 0.37 |
| **Mean** | **0.46** | **0.48** | **0.6** | **0.52** |

### B. Audio Modality

For the audio modality (Table III), performance improved slightly, achieving an average accuracy of 54%, precision of 55%, recall of 78%, and an F1-score of 63%. Notably, audio signals demonstrated greater consistency and higher recall, implying that vocal patterns might carry more robust indicators of deceptive speech compared to visual features.

TABLE III: Performance on Audio Modality

| Metric | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Fold 1 | 0.48 | 0.52 | 0.52 | 0.52 |
| Fold 2 | 0.59 | 0.58 | 0.84 | 0.68 |
| Fold 3 | 0.55 | 0.54 | 0.92 | 0.68 |
| Fold 4 | 0.57 | 0.58 | 0.68 | 0.62 |
| Fold 5 | 0.53 | 0.53 | 0.96 | 0.68 |
| **Mean** | **0.54** | **0.55** | **0.78** | **0.63** |

### C. Multimodal Fusion

Crucially, multimodal fusion significantly enhanced model performance, as illustrated in Table V. The fusion model achieved an average accuracy of 69%, precision of 74%, recall of 76%, and an F1-score of 72%. This improvement underscores the advantage of integrating multiple modalities, demonstrating how the complementary strengths of visual and audio features can be effectively leveraged to enhance predictive performance.

Table V further emphasizes the superior effectiveness of the multimodal fusion strategy compared to individual modalities. Specifically, multimodal integration provided a substantial increase in accuracy (+15 percentage points over audio, +23 percentage points over visual), precision (+19 points over audio, +26 points over visual), recall (+16 points over visual), and F1-score (+9 points over audio, +20 points over visual).

These results confirm that a multimodal approach is significantly beneficial for the deception detection task, reinforcing the hypothesis that combining different modalities yields richer representations, thus improving the overall accuracy and reliability of classification.

TABLE IV: Performance on Fusion

| Metric | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Fold 1 | 0.53 | 0.53 | 1.00 | 0.69 |
| Fold 2 | 0.74 | 0.78 | 0.72 | 0.75 |
| Fold 3 | 0.76 | 0.85 | 0.68 | 0.75 |
| Fold 4 | 0.72 | 0.70 | 0.84 | 0.76 |
| Fold 5 | 0.72 | 0.87 | 0.56 | 0.68 |
| **Mean** | **0.69** | **0.74** | **0.76** | **0.72** |

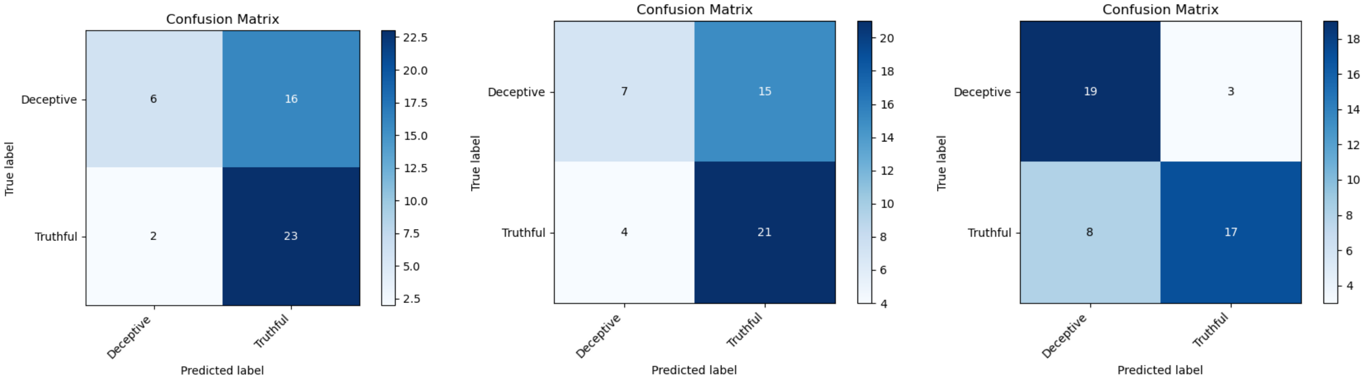
TABLE V: Comparison of Single-Modal and Fusion Performance

| Modality | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Visual | 0.46 | 0.48 | 0.6 | 0.52 |
| Audio | 0.54 | 0.55 | 0.78 | 0.63 |
| **Fusion** | **0.69** | **0.74** | **0.76** | **0.72** |

TABLE VI: Comparison of the proposed approach with those of Hsiao et al. [11] on the ATSFace dataset.

| Method | Accuracy | F1-score |
|---|---|---|
| Proposed approach | 0.76 | 0.68 |
| Hsiao et al. [12] | 0.79 | 0.79 |

As shown in Table VI, the proposed approach achieved an accuracy of 0.76 and an F1-score of 0.68, while the method by Hsiao et al. [11] obtained higher values in both metrics. Although the proposed model does not outperform the existing method, its performance is still competitive considering the constraints of the current setup. These results indicate that the approach holds promise and could benefit from further improvements, such as more extensive hyperparameter tuning, architectural enhancements, or training on larger and more diverse datasets.



(a) Video (fold 3)

(b) Audio (fold 2)

(c) Fusion (fold 3)

Fig. 3: Confusion matrices for the best fold of each modality and for the fusion.

## VI. Conclusion

Each model presented in this study was carefully designed and trained from scratch, deliberately avoiding the use of pretrained weights or external initialization strategies. Given the relatively limited size of the ATSFace dataset, a major challenge throughout the training and validation process was the recurring issue of overfitting. This constraint significantly limited the feasible complexity and depth of the neural network architectures, necessitating extensive experimentation with hyperparameter settings and the application of aggressive regularization techniques.

Despite these limitations, the proposed approach demonstrated promising performance, suggesting that with further optimization and access to larger datasets, it could serve as a viable and competitive solution for multimodal deception detection. Notably, while the method by Hsiao et al. [11] achieves superior performance in terms of accuracy and F1-score, it relies on a significantly more complex architecture. In contrast, the results obtained by the proposed model remain reasonably close, despite its simpler design and the constraints imposed by the limited dataset.

Future work will focus on enhancing the architecture and possibly incorporating transfer learning to further improve generalization capabilities.

## References

[1] A. Vrij, *Detecting Lies and Deceit: Pitfalls and Opportunities*. John Wiley & Sons, 2008.

[2] V. Pérez-Rosas, M. Abouelenien, A. Camacho, and R. Vijay, “Deception detection in the wild,” in *Proceedings of the 2015 IEEE International Conference on Multimedia and Expo (ICME)*, 2015.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[4] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, “Tensor fusion network for multimodal sentiment analysis,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114.

[5] S. Poria, E. Cambria, and D. Hazarika, “A review of affective computing: From unimodal analysis to multimodal fusion,” *Information Fusion*, vol. 37, pp. 98–125, 2017.

[6] S. Gupta *et al.*, “Bag-of-lies: A multimodal dataset for deception detection,” in *Proceedings of the 2019 Conference on Multimodal Deception Detection*, 2019.

[7] P. Ekman and W. V. Friesen, “Facial action coding system: A technique for the measurement of facial movement,” 1978.

[8] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[9] S. Feng, R. Banerjee, and Y. Choi, “Syntactic stylometry for deception detection,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, 2012, pp. 171–175.

[10] V. Karnati *et al.*, “Lienet: Multimodal deception detection using convolutional neural networks,” in *IEEE Transactions on Computational Social Systems*, 2021.

[11] S.-W. Hsiao and C.-Y. Sun, “Lora-like calibration for multimodal deception detection using atsface data,” Taipei, Taiwan, 2023, unpublished manuscript.