JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

MULTIMODAL DECEPTION DETECTION USING A FROZEN RESNET18 ENCODER AND ATTENTION-BASED FEATURE FUSION

Priya R. Ghongade, Prof.Vijaykumar M.P
Department of Computer Science and Engineering, Shreeyash College of Engineering & Technology, Chh. Sambhajinagar, India
Priyaghongade12@gmail.com

Abstract

Deception detection from human behaviour is difficult because no single cue, whether facial, vocal, lexical or behavioural, reliably indicates lying. This paper presents a multimodal framework that analyses four sources of evidence from a courtroom video clip: sampled video frames, speech descriptors, a TF-IDF transcript representation and 39 annotated gesture indicators. A frozen ImageNet-pretrained ResNet18 encodes the visual stream, small modality-specific networks project the remaining inputs into a shared 64-dimensional space, and a learned softmax attention module weights the four embeddings before a compact classifier predicts whether the speaker is truthful or deceptive. The learned part of the model has 96,614 trainable parameters. Experiments use the Real-life Trial dataset of 121 clips under a fixed, stratified 70/15/15 split, with all transformations fitted on training data only and a single test evaluation at a threshold of 0.50. Against four visual-only baselines (a scratch-trained CNN, EfficientNet-B0, ResNet18 and MobileNetV2), the proposed model reached 78.95% accuracy, 80.0% F1-score, 81.1% ROC-AUC and 88.4% PR-AUC on the 19-clip test set, and it was the best model on most metrics. A validation-only modality ablation indicated that audio contributed most among the added modalities. Because the test set is small, the results are treated as indicative rather than conclusive, and the paper discusses statistical uncertainty, clip-level splitting and possible annotation leakage.

Keywords— deception detection; multimodal learning; attention-based fusion; ResNet18; transfer learning; MFCC; TF-IDF; small-sample evaluation

I. INTRODUCTION

Deception, the deliberate attempt to make another person believe something the speaker knows to be false, is a common feature of human communication with serious consequences in criminal investigation, courtroom testimony, border security, insurance fraud and recruitment. Traditional judgements rely on human observation or on contact-based instruments such as the polygraph. Observers perform only modestly better than chance and are influenced by bias and by stereotyped beliefs about nonverbal cues [1], while physiological instruments measure arousal rather than deception itself.

Advances in computer vision and speech processing make it possible to treat deception as a pattern recognition problem. A recorded clip carries several parallel streams: facial expression and gaze in the video, spectral and prosodic structure in the audio, lexical content in the transcript and behavioural annotations of gestures. None of these streams is sufficient on its own, but a model that learns from all of them may capture weak and complementary evidence. Real-life courtroom recordings [2] provide a realistic, if small, testbed for such experiments.

This paper describes an end-to-end experimental framework of this kind. Its main design goal is stability under very small data. A pretrained visual encoder is kept frozen, the trainable part is limited to fewer than 100,000 parameters, and the evaluation protocol prevents information from validation or test subsets from entering any fitted transformation or model choice. The system is a research prototype for behavioural analysis and is not intended for legal or employment decisions.

II.RELATED WORK

A. Deception Datasets and Multimodal Evidence

Early studies often relied on laboratory data in which lies are role-played at low stakes, which limits ecological validity. Pérez-Rosas et al. [2] released a dataset of real courtroom clips labelled from trial outcomes. Later collections such as Bag-of-Lies [3] and Box of Lies [4] record deception in controlled or dialogue settings with several modalities. Surveys of multimodal

learning [5] describe the general benefit and the difficulty of combining heterogeneous channels, in particular the choice of fusion strategy.

B. Pretrained Visual Models and Transfer on Small Data

Transfer from large image corpora [6] is the usual route when labelled data are scarce. Residual networks [7], compact scaled networks [8] and inverted-residual mobile networks [9] are standard pretrained backbones. Fine-tuning all of their parameters on a few dozen clips is, however, prone to overfitting, which motivates freezing the encoder when data are limited.

C. Research Gap and Objectives

Three practical gaps motivate this work. First, fully fine-tuned networks can look strong on a small validation split and fail on a held-out split. Second, a single accuracy figure can hide one-sided classifiers that accept almost every or reject almost every clip. Third, the contribution of each added channel is seldom quantified under a protocol that keeps the test set untouched. The study pursues four objectives: (i) build a leakage-controlled pipeline for four modalities; (ii) test whether a frozen pretrained encoder with an attention-based fusion head is stable on 84 training clips; (iii) compare it with visual-only baselines under identical rules; and (iv) quantify the value of each modality on validation data only.

The contributions are: a reproducible, audited preprocessing pipeline; a light-weight attention-based fusion network with a frozen ResNet18 encoder; a same-protocol comparison with four visual baselines, including all results regardless of score; a validation-only modality ablation with analysis of the learned attention weights; and an evaluation that reports eight metrics and statistical uncertainty.

II. PROPOSED METHODOLOGY

A. Dataset and Splitting

The experiments use the Real-life Trial dataset [2], which contains 121 video clips from public courtroom proceedings: 61 deceptive and 60 truthful. Labels follow trial outcomes and are therefore proxies for verified lies. Each clip has a transcript and one row of 39 manually annotated binary gesture indicators covering facial expression, eyebrows, eyes, gaze, mouth, lips, head and hands. The manifest pairs every clip with its transcript and gesture row, and the folder-derived label is cross-checked against the annotation file, which reported no conflicts. The split is stratified by class with seed 42 into 84 training, 18 validation and 19 test clips, as summarised in Table I. The split is made at clip level because speaker identifiers are unavailable, so clips from one speaker may fall into different subsets.

TABLE I. Composition of training, validation and test subsets

Subset	Truthful	Deceptive	Total	Share (%)
Training	42	42	84	69.4
Validation	9	9	18	14.9
Test	9	10	19	15.7
All	60	61	121	100.0

B. Leakage-Controlled Preprocessing

Every statistic learned from data is computed on the training subset and then applied unchanged to validation and test data. This covers the TF-IDF vocabulary and inverse document frequencies, the median

imputation of gesture entries, and the standardisation of audio and gesture features. For a feature x with training mean μ and standard deviation σ, the standardised value is

$$z = \frac{x - \mu_{train}}{\sigma_{train}} \tag{1}$$

C. Feature Extraction

Video. Four frames are sampled at evenly spaced positions, converted to RGB, resized to 112 × 112 pixels and normalized with ImageNet statistics. A frozen ResNet18 with global average pooling (GAP) produces a 512-dimensional vector per frame [7], implemented in PyTorch [10], and the clip-level visual feature is their mean:

$$f_v = \left(\frac{1}{T}\right)\sum_{t=1}^{T} \varphi(X_t), T=4, f_v \in \mathbb{R}^{512} \tag{2}$$

Freezing the encoder, including its batch-normalization statistics, is a deliberate regularisation choice. Fine-tuning roughly 11 million parameters on 84 clips would be expected to overfit, whereas the design leaves about one hundred thousand parameters trainable. The resolution of 112 × 112 pixels was chosen to fit a 4 GB GPU; it discards fine facial detail and is acknowledged as a limitation.

Audio. The audio track is extracted with FFmpeg, converted to 16 kHz mono and truncated to 30 s. librosa

[11] yields 20 clip-level descriptors: duration, RMS energy (mean and standard deviation), zero-crossing rate, spectral centroid, bandwidth and roll-off, and the means of 13 MFCCs. All 121 clips produced non-zero features. Transcript. Each transcript is vectorised with TF-IDF (scikit-learn [12]) using the 300 most frequent uni- and bi-grams after removing English stop-words, fitted on training transcripts only. The weight of term t in document d is

$$tfidf(t,d) = tf(t,d) \times idf(t), idf(t) = \ln\left[\frac{1+n}{1+df(t)}\right] + 1 \quad (3)$$

where n is the number of training documents and df(t) the number of training documents containing t. Vectors are L2-normalised. Gestures. The 39 binary indicators are used directly, with training-median imputation and standardisation as in Eq. (1).

D. *Attention-Based Multimodal Fusion*

The proposed network has three stages: modality-specific projection, modality attention and weighted fusion followed by classification (Fig. 1). Each modality is mapped to a 64-dimensional embedding by its own small network. The visual vector (512 units) passes through a linear layer, layer normalization (LN), ReLU and dropout of 0.40. The audio branch (20 → 32 → 64), transcript branch (300 → 64 → 64, dropout 0.45) and gesture branch (39 → 32 → 64) follow the same pattern. Writing $h_m = g_m(f_m)$ for m in {v, a, t, g}:

$$h_m = g_m(f_m) \in \mathbb{R}^{64}, m \in \{v, a, t, g\} \quad (4)$$

The four embeddings are concatenated into H, and a two-layer attention network produces one score per modality, normalised by a softmax:

$$H = [h_v; h_a; h_t; h_g] \in \mathbb{R}^{256}, \alpha = softmax(W_2\, ReLU(W_1 H + b_1) + b_2) \in \mathbb{R}^{4} \quad (5)$$

where $\alpha = (\alpha_v, \alpha_a, \alpha_t, \alpha_g)$ are positive weights summing to one, W_1 maps 256 to 64 units and W_2 maps 64 to 4 units. Because the weights are computed for each clip, the model can emphasized different modalities for

different speakers. Each embedding is scaled by its weight and the results are concatenated, so each modality keeps its own slot in the fused vector:

$$f = [\alpha_v h_v; \alpha_a h_a; \alpha_t h_t; \alpha_g h_g] \in \mathbb{R}^{256} \quad (6)$$

The classifier C maps f through a 256 → 64 layer with LN, ReLU and dropout of 0.50, then a 64 → 2 layer. The probability of the deceptive class is thresholded at 0.50:

$$p = softmax(C(f)) \in \mathbb{R}^{2}, \hat{y} = 1[p \geq 0.50] \quad (7)$$

The attention module receives no supervision about which modality to trust; its only training signal is the classification loss, so its weights are treated as diagnostics rather than as proof of causal importance. The learned trainable part totals 96,614 parameters (Table II).

TABLE II. Trainable components of the proposed model

Component	Layers	Output	Trainable params
Visual encoder (frozen)	ResNet18 to GAP, mean over 4 frames	512	0 (≈11.2 M frozen)
Visual projection	512→64, LN, ReLU, Dropout 0.40	64	32,960
Audio branch	20→32→64, LN, ReLU, Dropout 0.40	64	2,976
Transcript branch	300→64→64, LN, ReLU, Dropout 0.45	64	23,680
Gesture branch	39→32→64, LN, ReLU, Dropout 0.40	64	3,584
Modality attention	256→64→4, ReLU, Dropout 0.40, Softmax	4	16,708
Classifier	256→64, LN, ReLU, Dropout 0.50, 64→2	2	16,706
Total trainable			96,614

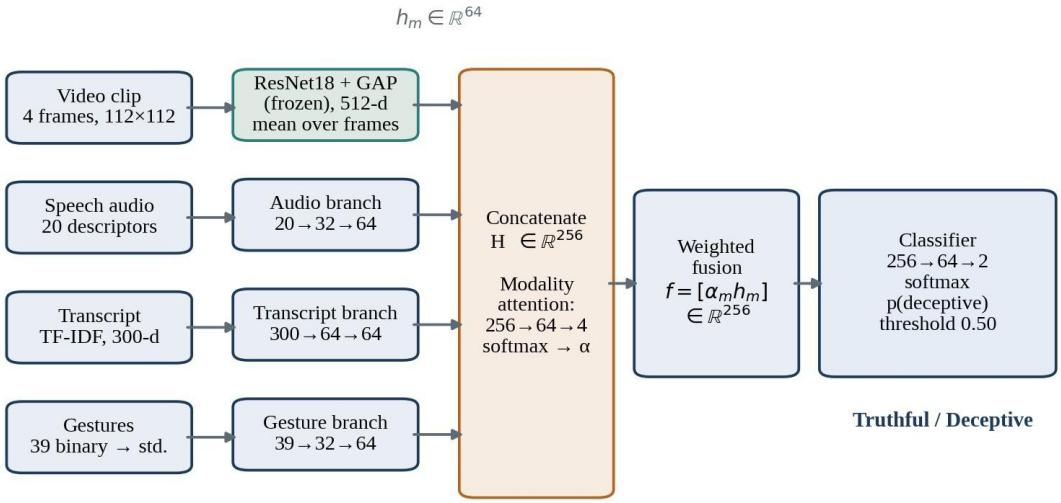

Fig. 1. Proposed attention-based multimodal fusion architecture. Dimensions follow Table II and Eqs. (2)–(7).

E. Visual-Only Baselines

Four baselines receive only the sampled frames and share the same training rules. A Base CNN with four convolutional blocks (32, 64, 128 and 256 filters, batch normalisation, ReLU) is trained from scratch. EfficientNet-B0 [8], ResNet18 [7] and MobileNetV2 [9] are ImageNet-pretrained and fully fine-tuned, each with a dropout (0.25) and linear head. Frame outputs are averaged to a clip-level prediction. The ResNet18 baseline uses the same backbone as the proposed model but not its audio, transcript or gesture inputs, so the comparison is at the level of modalities rather than architectures.

F. Training and Evaluation Protocol

All five models use AdamW [13] with learning rate 1×10^{-4} and weight decay 5×10^{-4}, batch size 2, at most 10 epochs, and early stopping with a patience of three epochs on validation loss. The loss is class-weighted cross-entropy:

$$L=-\left(\frac{1}{N}\right)\sum_{i=1}^{N} w_{y_i}\log p_{i,y_i}, w_c=\frac{N}{2N_c} \quad (8)$$

where N_c is the number of training clips of class c. The training set is exactly balanced, so both weights equal 1.0 in this study, but the mechanism is retained for imbalanced data. The parameters of the epoch with the lowest validation loss are restored, because validation accuracy on 18 clips changes in steps of 5.6 points. The test set is evaluated once per model at threshold 0.50, with no tuning after seeing test results. Figure 2 summarises the workflow.

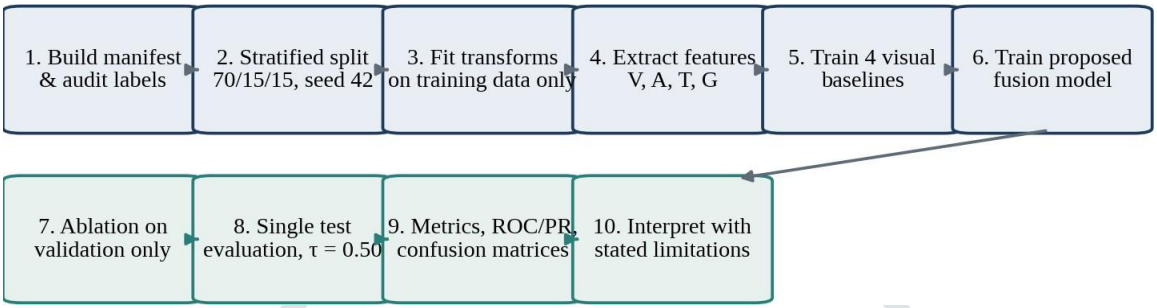

Fig. 2. Experimental workflow, from raw clips to the reported test results.

III. EXPERIMENTAL SETUP AND METRICS

Experiments ran in a Jupyter Notebook on a Windows laptop with an NVIDIA GeForce GTX 1650 Ti (4 GB VRAM). The software stack was Python 3.12, PyTorch and torchvision with pretrained ImageNet weights, OpenCV, FFmpeg through imageio-ffmpeg, librosa, scikit-learn, NumPy, pandas, and matplotlib. Python, NumPy and PyTorch seeds were fixed to 42 with deterministic cuDNN settings, so results are repeatable in the same environment. Audio extraction for all 121 clips took about three minutes and forty seconds.

Deceptive is the positive class. With TP, TN, FP and FN denoting counts of true and false positives and negatives at threshold 0.50, the metrics are

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}, Precision=\frac{TP}{TP+FP}, Recall=\frac{TP}{TP+FN} \quad (9)$$

$$F1=\frac{2\cdot Precision\cdot Recall}{Precision+Recall}, Specificity=\frac{TN}{TN+FP}, BalAcc=\frac{Recall+Specificity}{2} \quad (10)$$

Recall is sensitivity, specificity measures how many truthful clips are accepted, and balanced accuracy is robust to class imbalance. In addition, ROC-AUC (the probability that a random deceptive clip scores above a random truthful clip) and PR-AUC (average precision) are threshold-independent measures. Confidence intervals for accuracy use the Wilson score interval [14]. Because one clip changes test accuracy by about 5.3 points, differences of one or two clips are not treated as evidence of superiority.

IV. RESULTS AND DISCUSSION

A. Training Behaviour of the Baselines

Figure 3 plots validation loss and accuracy for all models, using the per-epoch training logs. The scratch-trained Base CNN learns slowly and unstably: its validation loss fluctuates between 0.68 and 0.85, and the selected epoch (5) has 44.4% validation accuracy. EfficientNet-B0 improves steadily, with validation loss falling from 0.574 to 0.287 at epoch 9 while validation accuracy stays between 78% and 83%. ResNet18 gives the best validation score (88.9%, lowest loss 0.254 at epoch 4), but its accuracy alternates between 72% and 89% from epoch to epoch. MobileNetV2 decreases smoothly, and its selected epoch is the last one

(77.8% validation accuracy), so a longer schedule may have helped. The proposed model has smooth curves, with training loss falling from 0.73 to 0.45 and validation loss from 0.67 to 0.49, and its best checkpoint is the final epoch, so it had not fully converged.

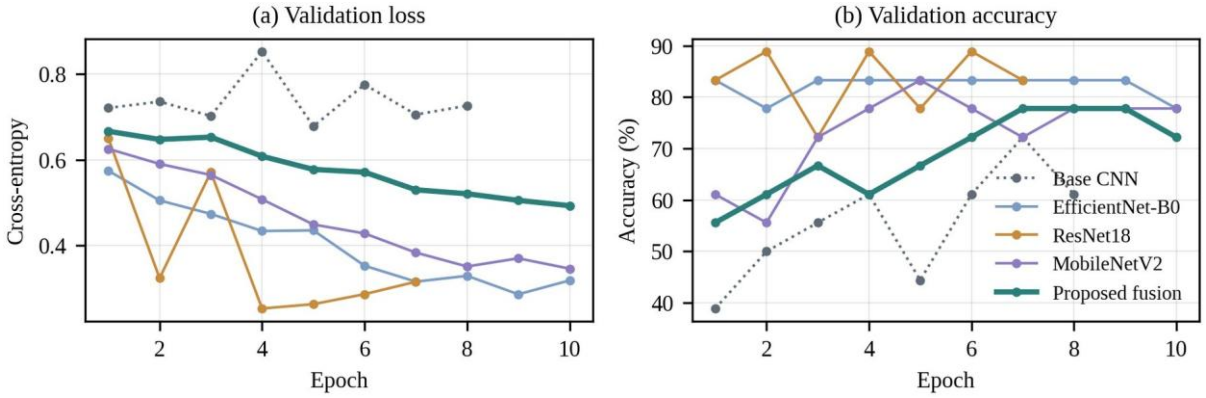

Fig. 3. Validation loss (a) and validation accuracy (b) of all models per epoch, from the per-epoch training logs of the experiments.

B. Held-Out Test Performance

Table III reports the test results at threshold 0.50 and Fig. 4 presents the core metrics graphically. The proposed model is best on accuracy, precision, F1-score, balanced accuracy, ROC-AUC and PR-AUC, and ties ResNet18 for the highest specificity (77.8%). Among the visual baselines, EfficientNet-B0 is strongest (73.7% accuracy, 76.2% F1-score), followed by MobileNetV2, the Base CNN and ResNet18.

TABLE III. Held-out test results of all models (threshold 0.50, n = 19), in %

Model	Acc.	Prec.	Recall	F1	Spec.	Bal. acc.	ROC-AUC	PR-AUC
Base CNN	57.89	56.25	90.00	69.23	22.22	56.11	70.00	79.87
EfficientNet-B0	73.68	72.73	80.00	76.19	66.67	73.33	70.00	75.15
ResNet18	47.37	50.00	20.00	28.57	77.78	48.89	76.67	76.43
MobileNetV2	63.16	63.64	70.00	66.67	55.56	62.78	61.11	71.04
Proposed fusion	78.95	80.00	80.00	80.00	77.78	78.89	81.11	88.44

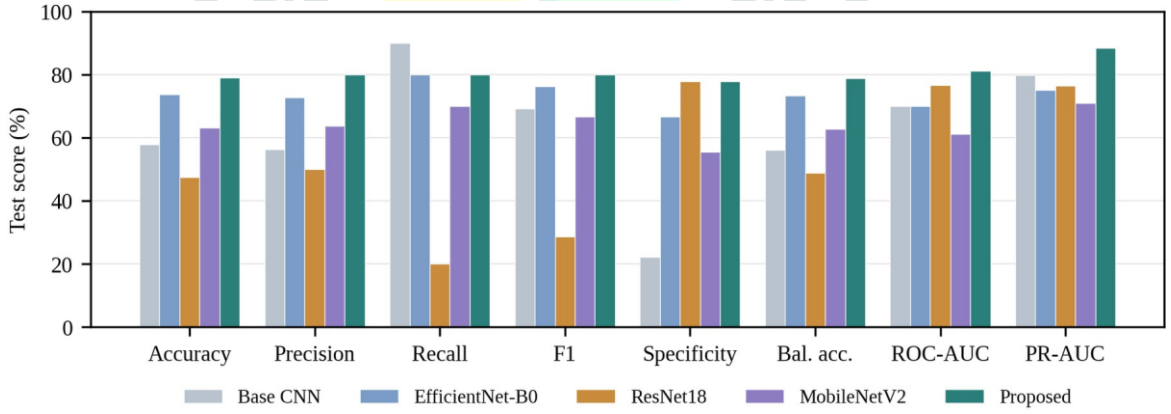

Fig. 4. Held-out test metrics of all models at threshold 0.50 (n = 19).

The baselines show characteristic one-sided behaviour (Fig. 5). The Base CNN finds 9 of 10 deceptive clips (recall 90%) but accepts only 2 of 9 truthful clips (specificity 22.2%). ResNet18 does the opposite, with recall 20% and specificity 77.8%. The ResNet18 result is the clearest case of validation misleading: the same model ranked first on validation and last on test. We attribute this to adaptation of a fully fine-tuned

network with about 11 million parameters to the idiosyncrasies of 84 training clips, and this observation motivated keeping the backbone frozen in the proposed model.

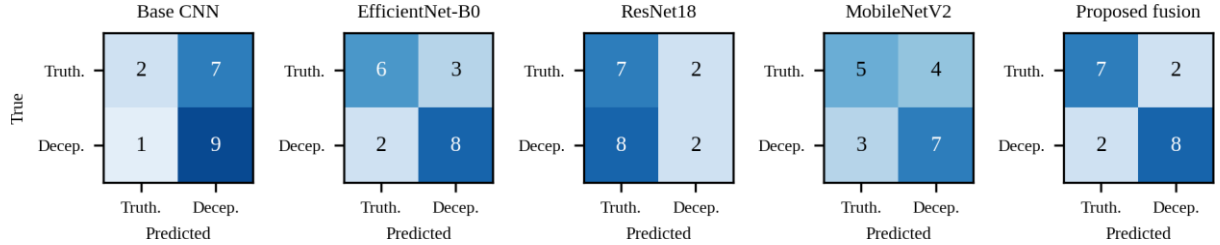

Fig. 5. Test confusion matrices of all models (rows: true class; columns: predicted class).

C. Proposed Model: Errors, Uncertainty and Attention

The proposed model classifies 7 of 9 truthful and 8 of 10 deceptive clips, with two false positives and two false negatives. Its errors are balanced between classes, so it does not trade one kind of mistake for the other. Its test accuracy corresponds to 15 of 19 clips correct, with a 95% Wilson interval of 56.7% to 91.5%. EfficientNet-B0 (14 of 19, interval 51.2% to 88.2%) has an interval that overlaps this one almost entirely, so the data do not support a statistically reliable ranking between the two. The ResNet18 interval (27.3% to 68.3%) lies mostly below that of the proposed model, and the proposed lower bound of 56.7% sits only slightly above chance. Averaged over the eight test metrics, the proposed model has the best mean rank (1.13), ahead of EfficientNet-B0 (2.50). A false-positive rate of 22% is nonetheless far too high for any decision about a person.

The learned attention weights on the validation set are visual 0.406, gestures 0.210, audio 0.202 and transcript 0.182 (Fig. 6b). The visual embedding receives the largest weight, while the three other modalities are close to the uniform value of 0.25. The attention module therefore mainly separates the visual stream from the rest and is not strongly selective, which is why the ablation is used as the more direct check on modality value.

D. Modality Ablation

Five variants of the fusion model were trained on the training split and compared on the validation set only, so that the test set was not used for this analysis (Table IV, Fig. 6a). Adding audio to video gave the largest improvement in accuracy (72.2% to 77.8%) and raised ROC-AUC from 76.5% to 85.2%. Adding gestures raised ROC-AUC to 81.5% and PR-AUC to 86.8% without changing accuracy. The transcript left accuracy and ROC-AUC unchanged and increased PR-AUC by only 0.3 points, which suggests that the sparse TF-IDF representation contributes little on its own. The all-modality model matches video plus audio on ROC-AUC (85.2%) but is one clip lower in accuracy.

TABLE IV. Validation results of the modality ablation (n = 18), in %

Variant	Acc.	Prec.	Recall	F1	Spec.	ROC-AUC	PR-AUC
Video only	72.22	70.00	77.78	73.68	66.67	76.54	83.07
Video + Audio	77.78	77.78	77.78	77.78	77.78	85.19	88.09
Video + Transcript	72.22	70.00	77.78	73.68	66.67	76.54	83.34
Video + Gestures	72.22	70.00	77.78	73.68	66.67	81.48	86.81
All modalities	72.22	70.00	77.78	73.68	66.67	85.19	87.75

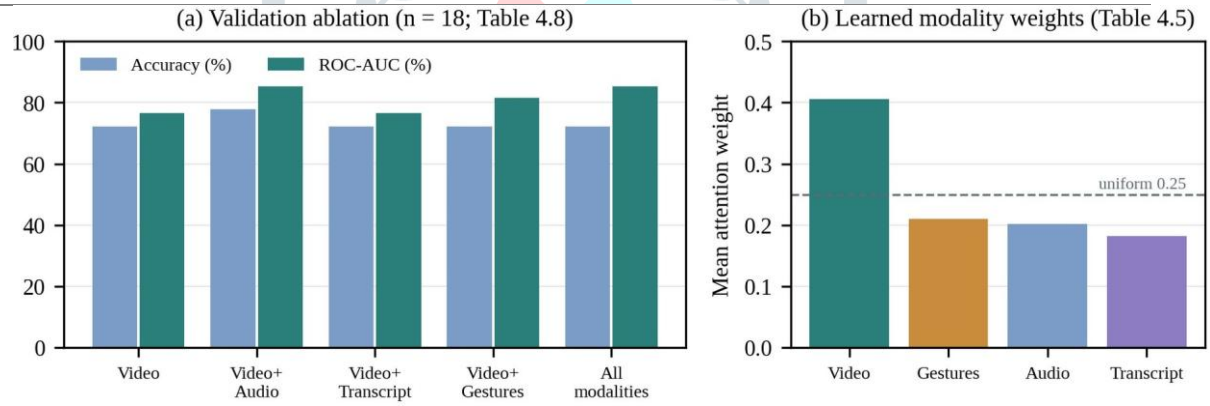

Fig. 6. Modality ablation on the validation set (a) and mean learned attention weights (b).

The video-only fusion variant (72.2%) is weaker on validation than the fine-tuned ResNet18 baseline (88.9%). This is not a contradiction: the ablation uses a frozen encoder with a small trainable head, whereas the baseline fine-tunes the full network. With only 18 validation clips, differences between variants are at the level of one or two clips, so the ablation suggests that audio is the most useful addition in this setup rather than providing a precise ranking of modalities.

E. Discussion and Threats to Validity

Four findings stand out. First, the proposed model achieved the best overall test performance and was the only model with balanced class-wise behaviour. Second, pretraining combined with regularisation matters on small data: the scratch-trained CNN was near chance on validation, and full fine-tuning overfitted. Third, validation rankings do not transfer reliably, so no ranking here should be regarded as definitive. Fourth, the comparison is not one of equal inputs, because the proposed model sees four modalities while the baselines see video only. The improvement may therefore partly reflect the added information rather than the fusion mechanism. A simple concatenation baseline was not evaluated and would clarify this. Four further threats apply. The statistical threat is the wide interval on 19 test clips. The selection threat is that one fixed split was used, and another seed could change the order of close models. Speaker dependence may make the task easier than a speaker-independent one. Construct validity is limited because the labels reflect trial outcomes and the gesture annotations were produced by people, so the model may learn annotation-specific regularities as well as behavioural cues.

The accuracy obtained here is of the same order as values reported for recent multimodal and specialised methods on realistic data, but those were obtained on larger datasets and different protocols. The results therefore do not support any claim of state-of-the-art performance. They do show that, under a strict protocol on a small real-life dataset, a light-weight attention-based fusion of a frozen pretrained encoder with simple audio, text and gesture features is a stable and competitive design, while also making clear how large the uncertainty is with 19 test clips.

V. CONCLUSION AND FUTURE SCOPE

This paper presented a multimodal deception detection framework that analyses video, speech, transcript and annotated gestures from a courtroom clip. A frozen ResNet18 encodes sampled frames, compact branches project the other modalities into a shared 64-dimensional space, and a learned attention module weights the four embeddings before a small classifier. The learned part has 96,614 trainable parameters. Under a leakage-controlled protocol the model achieved 78.95% accuracy, 80.0% F1-score, 81.1% ROC-AUC and 88.4% PR-AUC on 19 test clips, and it was the best of five models on most metrics. A validation-only ablation indicated that audio was the most useful addition to video, and that the sparse transcript representation contributed little.

The study is limited by its size: the confidence intervals overlap, validation and test rankings disagree, and attention weights are only weakly selective. Future work should evaluate speaker-independent splits with repeated seeds, obtain larger and more diverse data and test cross-dataset generalisation. Richer encoders (learned speech representations, transformer text embeddings and temporal video models) could replace the summary features, and automatically extracted facial action units and head-pose features could replace manual gesture annotation, removing the leakage risk. Stronger fusion baselines such as concatenation, gated and cross-modal attention would clarify what the attention module contributes. Explainability and probability calibration would make the output more transparent.

Ethical use is essential. A prediction is a statistical estimate that behaviour resembles patterns in one small labelled dataset; it is not a measurement of lying, and anxiety, fatigue, cultural differences and illness can all produce the same patterns. The system must not be used to make or support legal, employment or disciplinary decisions. Any real application would require independent validation, bias testing across demographic groups, informed consent and a responsible human decision-maker.

REFERENCES

[1] A. Vrij, M. Hartwig, and P. A. Granhag, “Reading lies: Nonverbal communication and deception,” Annu. Rev. Psychol., vol. 70, pp. 295–317, 2019.

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

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

[4] F. Soldner, V. Pérez-Rosas, and R. Mihalcea, “Box of Lies: Multimodal deception detection in dialogues,” in Proc. NAACL-HLT, 2019, pp. 1768–1777.

[5] 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.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.

[7] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE CVPR, 2016, pp. 770–778.

[8] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. Int. Conf. Machine Learning (ICML), 2019, pp. 6105–6114.

[9] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proc. IEEE CVPR, 2018, pp. 4510–4520.

[10] A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” in Advances in Neural Information Processing Systems (NeurIPS), 2019, pp. 8024–8035.

[11] B. McFee et al., “librosa: Audio and music signal analysis in Python,” in Proc. 14th Python in Science Conf. (SciPy), 2015, pp. 18–24.

[12] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[13] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in Proc. Int. Conf. Learning Representations (ICLR), 2019.

E. B. Wilson, “Probable inference, the law of succession, and statistical inference,” J. Amer. Statist. Assoc., vol. 22, no. 158, pp. 209–