Lightweight Multimodal Emotion Recognition with Intermediate Fusion for Edge Deployment

Donny Halim[1], Fitra Abdurrachman Bachtiar[1*], and Tirana Noor Fatyanosa[1]

Department of Informatics Engineering, Faculty of Computer Science,
Universitas Brawijaya,
Malang, East Java, Indonesia
fitra.bachtiar@ub.ac.id

Abstract. Multimodal emotion recognition benefits affective computing applications such as companion robots, driver monitoring, and assistive healthcare, yet most audio-visual systems rely on heavy backbones and static fusion mechanisms that are unsuitable for edge devices or cannot adapt to input-dependent modality reliability. This paper proposes a lightweight framework using EfficientNetV2-B0 for facial features and a compact 1D-CNN for prosodic audio features (MFCC + Delta + Delta-Delta), integrated through a Gated Multimodal Unit (GMU) that adaptively weighs each modality via a learned gate. It targets a Raspberry Pi 4B with constraints of <12M parameters, <2 GFLOPs, <50 MB model size, and <100 ms inference time. Evaluation on RAVDESS (1,440 recordings, 8 emotion classes) uses actor-independent 5-fold group cross-validation, comparing GMU with static Concatenation and Cross-Attention fusion under identical protocols, together with masking, ablation, and roofline analyses. GMU achieves 94.17% (±3.94%) mean cross-validation accuracy and held-out test accuracy of 77.08% (single model) and 79.17% (5-fold ensemble), with 7.75M parameters, 1.41 GFLOPs, 30.21 MB model size, and 15.22 ms inference latency, meeting all deployment constraints, unlike Cross-Attention (78.33% test accuracy, 19.69M parameters). Masking shows audio adds 9.58 percentage points over the visual-only configuration, and roofline analysis indicates a compute-bound regime consistent with Raspberry Pi 4B characteristics. GMU offers the best accuracy-efficiency trade-off among constraint-compliant fusion methods and is recommended as the default intermediate fusion strategy for real-time, resource-constrained emotion recognition; Cross-Attention may suit settings without edge constraints, at the cost of a larger memory and parameter footprint.

Keywords: Multimodal Emotion Recognition · Gated Multimodal Unit · Lightweight Architecture · Intermediate Fusion · Edge Deployment

1 Introduction

Emotion recognition enables machines to perceive human affective states, supporting applications such as companion robots, driver monitoring, adaptive learning, and assistive healthcare [5, 12, 14]. Because emotions are expressed jointly through facial expressions and vocal characteristics, multimodal audio-visual approaches consistently outperform

© The Author(s) 2026
M. Murtadha et al. (eds.), *Proceedings of the 2026 11th International Conference on Sustainable Information Engineering and Technology (SIET 2026)*, Atlantis Highlights in Computer Sciences 26,
https://doi.org/10.2991/978-94-6239-828-3_16

unimodal counterparts, particularly under degraded or partial sensory conditions [2, 6]. However, state-of-the-art multimodal systems typically rely on deep Transformers or high-capacity CNN backbones, demanding computational resources unsuitable for resource-constrained edge devices [13, 7].

Among fusion strategies, intermediate fusion offers a favorable trade-off: unlike early fusion, it avoids strict temporal synchronization requirements, and unlike late (decision-level) fusion, it preserves cross-modal correlations by combining modality embeddings before the final classification stage [4, 11]. Static intermediate fusion methods such as feature concatenation are parameter-free but cannot adapt the relative contribution of each modality to input-dependent signal quality. The Gated Multimodal Unit (GMU) [1] addresses this limitation through a learned gating mechanism that adaptively weighs modality embeddings, offering a lightweight alternative to attention-based fusion.

We investigate four research questions: whether GMU improves accuracy over static/attention fusion; the complementary contributions of each modality; satisfaction of deployment constraints for Raspberry Pi 4B; and influential training/architectural configurations. We use EfficientNetV2-B0 [17] for faces and a compact 1D-CNN for prosodic audio, with masking, roofline, and ablation analyses under identical protocols. All evaluations use actor-grouped 5-fold cross-validation on RAVDESS [8] to ensure statistically reliable, subject-independent estimates.

The paper is organized as follows. Section 2 reviews related work. Section 3 presents the methodology. Section 4 describes experiments and results. Section 5 discusses findings. Section 6 concludes and provides recommendations.

2 Related Work

Prior multimodal emotion recognition studies confirm that combining facial expression and speech yields the most robust performance among modality pairs [11], but achieving this typically requires architectures exceeding edge-device budgets. Luna-Jiménez et al. [10] achieved 80.08% accuracy on RAVDESS through late fusion of CNN-14 and STN features using 81M parameters. Chumachenko et al. [6] proposed self-attention fusion with modality dropout, achieving 81.58% accuracy, and demonstrated that attention-based intermediate fusion improves robustness to missing modalities. Ben Letaifa and Boli [3] introduced transformer-based cross-attention fusion with an EfficientNetV2-B0 backbone, achieving 88.33% accuracy with 14.9M parameters, while their concatenation-based variant achieves 87.29% accuracy with 13.1M parameters – illustrating the accuracy-efficiency trade-off between static and attention-based fusion that this work also investigates.

On the lightweight side, Radoi and Cioroiu [13] proposed uncertainty-based multimodal learning using 2.7M parameters, reaching 76.30% accuracy. Lu et al. [9] reported 97.56% accuracy with only 0.92M parameters using GhostNet and 1D-CNN with decision-level fusion, though under a different (non-actor-independent) evaluation protocol that limits direct comparability. EfficientNetV2-B0 offers a favorable accuracy-complexity trade-off for visual feature extraction at 7.1M parameters and low FLOPs [17], while compact 1D-CNNs on prosodic features provide low-latency alternatives to spectrogram-based 2D architectures [15]. Boulahia et al. [4] concluded intermediate

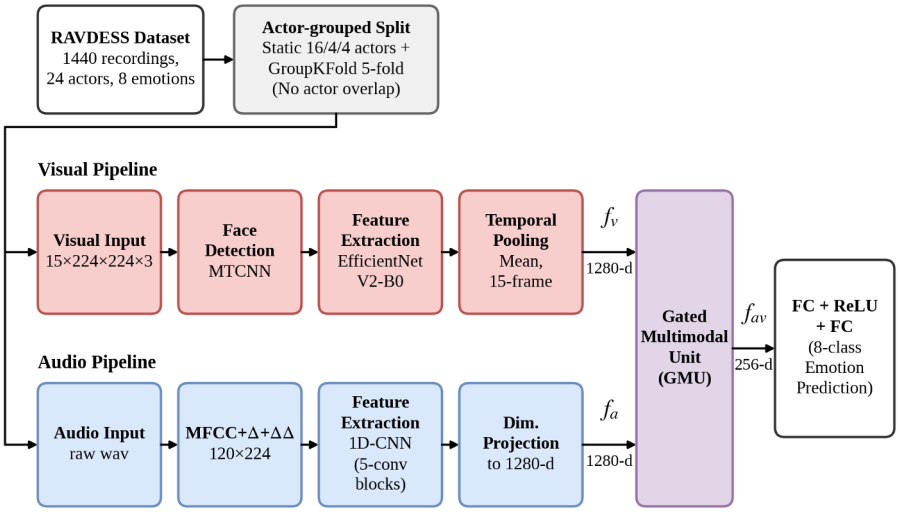


Fig. 1. Proposed architecture: visual pipeline with per-frame EfficientNetV2-B0 features aggregated via mean pooling; audio pipeline with 1D-CNN over MFCC+Delta+Delta-Delta; Gated Multimodal Unit (GMU) intermediate fusion; and classifier.

fusion achieves the best performance-complexity balance, and Arevalo et al. [1] introduced GMU for adaptive weighting. We position GMU as a middle ground between static concatenation and attention: adaptive like attention, yet with substantially lower parameter overhead for edge constraints.

3 Proposed Methodology

3.1 System Overview

Fig. 1 illustrates the proposed architecture: two parallel feature-extraction pipelines for visual and audio modalities, a GMU-based intermediate fusion module, and a classification head.

3.2 Dataset and Protocol

RAVDESS [8] consists of 1,440 recordings from 24 actors (12 male, 12 female) covering eight emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. A static subject-wise split (16 training / 4 validation / 4 testing actors, i.e., 960/240/240 samples) is used to report a held-out test set with no actor overlap. To obtain a more statistically reliable estimate of generalization, multimodal models are additionally evaluated with actor-grouped 5-fold cross-validation (GroupKFold), ensuring no actor appears in both training and validation folds of any split. The 5-fold cross-validation serves as an internal stability check, whereas the static held-out split is preserved as the primary test benchmark to maintain comparability with prior studies that use this fixed partition.

3.3 Visual and Audio Feature Extraction

MTCNN is used for face detection and alignment; each video is sampled into 15 uniformly distributed frames, cropped and resized to 224×224×3. EfficientNetV2-B0 [17], pretrained on ImageNet, extracts a 1280-dimensional embedding per frame; frame-level embeddings are aggregated via mean pooling into a video-level visual feature vector $\mathbf{f}_v \in \mathbb{R}^{1280}$.

Each audio track is converted into Mel-Frequency Cepstral Coefficients (MFCC) together with first and second-order derivatives (Delta and Delta-Delta), yielding a 120-channel, 224-frame representation. This representation is processed by a compact 1D-CNN with five convolutional blocks (channel widths 160-160-128-96-64) followed by global average pooling and a linear projection, producing an audio embedding $\mathbf{f}_a \in \mathbb{R}^{1280}$, dimensionally matched to the visual embedding for fusion.

3.4 Gated Multimodal Unit (GMU) Fusion

Unlike static concatenation, GMU [1] learns an input-dependent gate z that adaptively controls the relative contribution of each modality. Each modality embedding is first projected into a shared fusion space of dimension $d = 256$:

$$\mathbf{h}_a = \tanh(W_a \mathbf{f}_a), \qquad \mathbf{h}_v = \tanh(W_v \mathbf{f}_v) \tag{1}$$

$$z = \sigma\big(W_z [\mathbf{f}_a; \mathbf{f}_v]\big) \tag{2}$$

$$\mathbf{f}_{av} = z \odot \mathbf{h}_a + (1 - z) \odot \mathbf{h}_v \tag{3}$$

where $W_a, W_v \in \mathbb{R}^{256\times1280}$ project each modality embedding independently, $W_z \in \mathbb{R}^{256\times2560}$ computes the gate from the concatenation of the two *unprojected* 1280-d embeddings, σ is the sigmoid function, and $\odot$ denotes element-wise multiplication. The fused representation $\mathbf{f}_{av} \in \mathbb{R}^{256}$ passes through a two-layer fully connected classification head (256→128→8) to produce the emotion prediction. Fusion dimension $d = 256$ is the baseline configuration examined in the ablation study (Section 4.4).

3.5 Training Strategy and Deployment Constraints

Models are trained with cross-entropy loss using label smoothing ($\epsilon = 0.1$) [16] and hard modality dropout ($p = 0.3$) [6], using AdamW with cosine annealing learning-rate scheduling and early stopping. Feature caching accelerates multimodal training. Three fusion variants are trained and compared under identical conditions: static Concatenation (baseline), GMU (proposed), and Cross-Attention (upper-bound accuracy reference). Table 1 presents deployment constraints derived from Raspberry Pi 4B specifications, used to evaluate the practical feasibility of each fusion variant.

4 Experiments and Results

4.1 Experimental Setup

All experiments use PyTorch with an NVIDIA GPU for training; inference metrics are measured per sample. Evaluation metrics include accuracy, weighted and macro F1-score, parameter count, FLOPs, model size, and inference latency.

Table 1. Lightweight Deployment Constraints (Raspberry Pi 4B Target)

Metric	Target
Total Parameters	< 12 M
Computational Complexity	< 2 GFLOPs
Model Size	< 50 MB
Inference Time	< 100 ms

Table 2. Unimodal Performance on RAVDESS Test Set

Model	Acc (%)	Best Val (%)	F1-Macro	Params (M)	Size (MB)	GFLOPs
Audio (1D-CNN)	52.92	73.44	0.5104	1.20	4.61	0.117
Visual (EffNetV2-B0)	68.33	94.27	0.6814	6.52	25.44	1.30

4.2 Unimodal Baseline Results

Table 2 presents unimodal performance. The visual model substantially outperforms the audio model (68.33% vs. 52.92% test accuracy), indicating that facial expression is a stronger emotion indicator than vocal prosody on RAVDESS, consistent with the dataset's acted, highly expressive facial recordings [8].

4.3 Multimodal Fusion Results

Table 3 compares the three fusion variants; all multimodal models use actor-grouped 5-fold cross-validation to estimate stability, followed by held-out test evaluation.

GMU achieves the highest constraint-compliant test accuracy (77.08% for the best single model and 79.17% for the 5-fold ensemble) and F1-macro (0.8001), outperforming Concatenation (77.92%) with only 0.66M extra parameters. Cross-Attention yields 78.33% but fails the parameter and size budgets (19.69M, 76.78 MB).

The confusion matrix (Fig. 2) shows highest per-class accuracy for Calm (93.75%) and Angry (90.63%), while Fearful (62.50%) and Sad (65.63%) are lower, consistent with prior work [6, 10]. Fearful is most often confused with Angry, and Sad with Calm, reflecting overlapping prosodic/facial cues. GMU notably improves ambiguity resolution over the unimodal visual matrix, confirming effective leverage of complementary audio cues.

Masking Evaluation. Table 4 reports GMU accuracy when one modality is disabled at inference time. Audio contributes a 9.58 percentage-point gain over the visual-only

Table 3. Multimodal Fusion Comparison on RAVDESS Test Set

Model	5-Fold CV (%)	Test Acc (%)	F1-Macro	Params (M)	GFLOPs	Size (MB)	Status
Concatenation	94.17 ±3.86	77.92	0.7863	7.09	1.41	27.65	Pass
GMU (best single)	94.17 ±3.94	77.08	0.7745	7.75	1.41	30.21	Pass
GMU (5-fold ensemble)	**94.17 ±3.94**	**79.17**	**0.8001**	**7.75**	**1.41**	**30.21**	**Pass**
Cross-Attention	95.00 ±3.35	78.33	0.7871	19.69	1.44	76.78	Fail

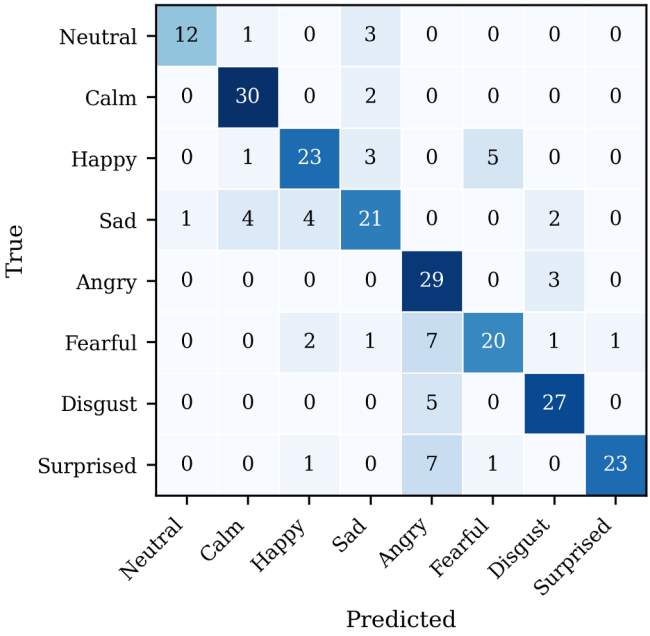


Fig. 2. Confusion matrix of the GMU model (best single model) on the RAVDESS test set.

Table 4. GMU Masking Evaluation

Mode	Accuracy (%)	F1-Macro
Audio + Visual	77.08	0.7745
Visual Only	67.50	0.6602
Audio Only	51.25	0.4842

configuration (77.08% vs. 67.50%), confirming that although audio is comparatively weak unimodally, it provides complementary information that GMU exploits effectively.

4.4 Ablation Study

Table 5 presents GMU ablation results on the validation set, varying label smoothing, modality dropout, and fusion dimension while holding other settings at baseline (fusion dimension 256). Fusion dimension has the largest effect: increasing it from 128 to 512 improves accuracy by 4.17 points, though dimension 256 is retained as the primary configuration to preserve the lightweight parameter budget. Modality dropout slightly reduces validation accuracy (baseline vs. no-dropout: −1.67 points) due to the visual modality's dominance on RAVDESS, though it remains beneficial for robustness to missing modalities at inference, as confirmed by the masking evaluation above.

Table 5. GMU Ablation Study (Validation Set)

Variant	Val Acc (%)	F1-Macro	Δ vs. Baseline
Baseline (LS + MD + dim 256)	73.75	0.7246	–
No Label Smoothing	73.33	0.7295	−0.42
No Modality Dropout	75.42	0.7426	+1.67
Fusion Dim 128	72.50	0.7194	−1.25
Fusion Dim 512	76.67	0.7639	+2.92

Table 6. Roofline Analysis of Multimodal Models

Model	Arithmetic Intensity	Ridge Point	Regime
Concatenation	51.08	2.34	Compute-Bound
GMU	46.88	2.34	Compute-Bound
Cross-Attention	50.95	2.34	Compute-Bound

4.5 Efficiency, Roofline, and State-of-the-Art Comparison

The roofline ridge point is computed assuming a Raspberry Pi 4B with peak performance of approximately 13.5 GFLOPS and memory bandwidth of approximately 5.8 GB/s, yielding a ridge point of 2.34 FLOPs/byte. Table 6 presents the roofline analysis, computed as the ratio of FLOPs to memory bytes moved (arithmetic intensity), against this ridge point. All three multimodal variants operate well above the ridge point, confirming a compute-bound regime in which performance is limited by processor throughput rather than memory bandwidth – consistent with the Raspberry Pi 4B's constrained compute-to-bandwidth ratio, and supporting the suitability of the proposed models for this target device.

Table 7 compares GMU against related work on RAVDESS. Lu et al. [9] report the highest accuracy (97.56%) with an ultra-compact architecture, but rely on decision-level fusion under a different evaluation protocol, which limits direct comparability and does not exploit semantic correlation between facial and vocal cues during representation learning. Ben Letaifa and Boli [3] and Chumachenko et al. [6] achieve higher accuracy through attention-based fusion, but at parameter counts (14.9M, and architecturally comparable) that exceed or approach the edge budget targeted here. Against Luna-Jiménez et al. [10], GMU trades 0.91 accuracy points for a 10× reduction in parameters (7.75M vs. 81M). Compared to Radoi and Cioroiu's uncertainty-based lightweight fusion [13], GMU improves accuracy by 2.87 points while remaining within the lightweight budget, indicating adaptive intermediate fusion provides tangible benefit over uncertainty-weighted fusion at comparable model scale.

5 Discussion

GMU adaptively weights modalities via its learned gate, yielding a consistent edge over static Concatenation in the ensemble setting (79.17% vs. 77.92%) with only 0.66M added parameters. Against Cross-Attention, GMU trades 1.25 accuracy points in the single-model setting but saves 2.5× parameters and memory; in the ensemble setting,

Table 7. Comparison with State-of-the-Art on RAVDESS

Method	Acc (%)	Params (M)	Fusion	Note
Luna-Jiménez et al. (2021) [10]	80.08	81.00	Late	Heavyweight
Chumachenko et al. (2022) [6]	**81.58**	–	Self-Attention	–
Radoi & Cioroiu (2024) [13]	76.30	2.70	Uncertainty-based	Lightweight
Lu et al. (2026) [9]	97.56	0.92	Decision-level	Different protocol
Ben Letaifa & Boli (2026) [3]	88.33	14.90	Cross-Attention	Heavyweight
Ours (GMU, ensemble)	**79.17**	**7.75**	**Intermediate GMU**	**Meets edge constraint**

GMU surpasses Cross-Attention (79.17% vs. 78.33%), making it preferable under edge constraints. Visual features dominate RAVDESS (acted, expressive corpus), yet masking shows audio adds a 9.58-point gain, confirming complementarity [2]. Limitations include evaluation on a single acted dataset, the 1D-CNN's moderate capacity versus pretrained encoders, and reliance on simulated (not physical) profiling. Future work includes in-the-wild validation, hardware profiling, temporal attention pooling, and higher fusion dimensions.

To support independent verification and extension of this work, the full implementation is available at https://github.com/donnyhalim/multimodal-emotion-recognition.

6 Conclusion and Recommendation

This paper proposed a lightweight multimodal emotion recognition framework combining EfficientNetV2-B0 visual features and 1D-CNN audio features through GMU intermediate fusion, targeting Raspberry Pi 4B. Using actor-grouped 5-fold cross-validation on RAVDESS, GMU achieved 94.17% mean CV accuracy and held-out test accuracies of 77.08% (single) and 79.17% (ensemble), with 7.75M parameters, 1.41 GFLOPs, 30.21 MB size, and 15.22 ms latency – meeting all constraints (<12M, <2 GFLOPs, <50 MB, <100 ms). GMU outperformed static Concatenation and remained lighter than Cross-Attention, which failed constraints. Masking and roofline analyses confirmed complementarity and compute-bound suitability. Thus, GMU improves accuracy over static fusion while staying lightweight, modalities are complementary (visual-dominant), constraints are satisfied, and fusion dimension is the most influential ablation factor.

GMU-based intermediate fusion is recommended for edge deployment when <1M extra parameters are acceptable, offering adaptive robustness. Cross-attention is reserved for scenarios without strict budgets. Future extensions should prioritize in-the-wild validation and physical profiling on Raspberry Pi 4B hardware.

Acknowledgments. Part of this research was funded by the Artificial Intelligence Applied Research Funding Program at Brawijaya University for the 2025 academic year under contract No. 02340.14/UN10.A0501/B/PT.01.03/2025.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arevalo, J., Solorio, T., Montes-y Gómez, M., González, F.A.: Gated multimodal units for information fusion. arXiv preprint arXiv:1702.01992 (2017)
2. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 423–443 (2019). https://doi.org/10.1109/TPAMI.2018.2798607
3. Ben Letaifa, L., Bohi, A.: Efficient audiovisual fusion architectures for emotion recognition. Neural Computing and Applications (2026). https://doi.org/10.1007/s00521-026-12056-5
4. Boulahia, S.Y., Amamra, A., Madi, M.R., Daikh, S.: Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. Machine Vision and Applications **32**, 121 (2021). https://doi.org/10.1007/s00138-021-01249-8
5. Calvo, R.A., D'Mello, S.: Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Transactions on Affective Computing **1**(1), 18–37 (2010). https://doi.org/10.1109/T-AFFC.2010.1
6. Chumachenko, K., Iosifidis, A., Gabbouj, M.: Self-attention fusion for audiovisual emotion recognition with incomplete data. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 2822–2828. IEEE (2022). https://doi.org/10.1109/ICPR56361.2022.9956447
7. Gonçalves, J., et al.: Efficient deep learning models for real-time emotion recognition on edge devices. IEEE Access **12**, 11345–11358 (2024). https://doi.org/10.1109/ACCESS.2024.3351234
8. Livingstone, S.R., Russo, F.A.: The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE **13**(5), e0196391 (2018). https://doi.org/10.1371/journal.pone.0196391
9. Lu, C.K., Lu, C.W., Lin, G.B.: Lightweight multimodal emotion recognition for companion robots: A deep learning framework integrating facial and speech features. IEEE Transactions on Affective Computing **17**(1), 412–425 (2026). https://doi.org/10.1109/TAFFC.2025.3541234
10. Luna-Jiménez, C., Griol, D., Callejas, Z., Kleinlein, R., Montero, J.M., Fernández-Martínez, F.: Multimodal emotion recognition on RAVDESS dataset using transfer learning. Sensors **21**(22), 7665 (2021). https://doi.org/10.3390/s21227665
11. Pan, B., Hirota, K., Jia, Z., Dai, Y.: A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. Neurocomputing **561**, 126866 (2023). https://doi.org/10.1016/j.neucom.2023.126866
12. Picard, R.W.: Affective Computing. MIT Press, Cambridge, MA (1997)
13. Radoi, A., Cioroiu, G.: Uncertainty-based learning of a lightweight model for multimodal emotion recognition. IEEE Access **12**, 120374–120390 (2024). https://doi.org/10.1109/ACCESS.2024.3450674
14. Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication **53**(9-10), 1062–1087 (2011). https://doi.org/10.1016/j.specom.2011.01.011
15. Singh, Y.B., Goel, S.: A lightweight 2D CNN based approach for speaker-independent emotion recognition from speech with new Indian emotional speech corpora. Multimedia Tools and Applications **82**, 23055–23073 (2023). https://doi.org/10.1007/s11042-023-14678-6

16. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the Inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826. IEEE (2016). https://doi.org/10.1109/CVPR.2016.308
17. Tan, M., Le, Q.: EfficientNetV2: Smaller models and faster training. In: Proceedings of the 38th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 139, pp. 10096–10106. PMLR (2021)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

