TERT-Ensemble: A Two-Stage Fusion Approach for Tri-modal Emotion Recognition

Penumuchu Nihith
CINTEL, Student
SRM Institute of Science & Technology
Kattankulathur, Tamil Nadu, India
ns2995@srmist.edu.in

M Lingeshwaran
CINTEL, Student
SRM Institute of Science & Technology
Kattankulathur, Tamil Nadu, India
lv7483@srmist.edu.in

Dr. Navneet Nayan
CINTEL, Assistant Professor
SRM Institute of Science & Technology
Kattankulathur, Tamil Nadu, India
navneetn@srmist.edu.in

***Abstract*—Automated emotion recognition is essential for advancing human-computer interaction and affective computing. While unimodal systems using image, text, or audio offer valuable insights, they often struggle with the inherent ambiguity and complexity of human emotional expression. Multimodal approaches promise enhanced robustness by integrating complementary information from diverse sources. This paper introduces 'TERT-Ensemble', a technique employing a two-stage fusion process for tri-modal emotion recognition. Initially, we establish strong unimodal predictors by systematically evaluating diverse deep learning architectures (CNNs, ViT, Transformers, LSTM) for each modality (image, text, audio) on combined benchmark datasets (CK+, FER-2013, RAF-DB; Twitter datasets; CREMA-D, RAVDESS, SAVEE). Optimal performance within each modality was achieved through intra-modal ensembles, yielding weighted F1-scores of 77.98% (Image), 73.69% (Text), and 66.43% (Audio) on their respective test sets. Subsequently, a tri-modal late fusion model was implemented, combining the outputs of these intra-modal ensembles via weighted probability averaging. Evaluated on a simulated tri-modal test set designed to ensure congruent emotional labels across modalities, this final fusion model achieved high performance, with an accuracy of 95.56% and a weighted F1-score of 0.96. We detail the data preprocessing, model architectures, training protocols, and the two-stage fusion strategy implemented within Kaggle notebooks. An interactive demonstration interface showcasing the unimodal components is also presented. While acknowledging the limitations inherent in using simulated data for the final fusion evaluation, these results validate the effectiveness of the intra-modal ensembles and highlight the significant potential of the proposed staged fusion approach for robust tri-modal emotion recognition.**

***Index Terms*—Emotion Recognition, Multimodal Learning, Late Fusion, Ensemble Learning, Deep Learning, Affective Computing, Image Recognition, Text Classification, Speech Emotion Recognition, Convolutional Neural Networks (CNN), Vision Transformer (ViT), BERT, DeBERTa, LSTM, EfficientNet, Gradio.**

I. Introduction

The automatic recognition of human emotions is a cornerstone for progress in diverse fields, including mental health diagnostics, driver safety systems, personalized content delivery, and the creation of more natural and empathetic human-computer interaction (HCI) systems [1]. Human emotions are inherently multimodal phenomena, conveyed concurrently through facial expressions, vocal prosody, linguistic content, and other physiological cues [2]. Traditional computational systems often focus on analyzing these cues in isolation, leading to specialized domains like Facial Expression Recognition (FER) [3], text-based sentiment/emotion analysis [4], and Speech Emotion Recognition (SER) [5].

However, these unimodal approaches inherently struggle to capture the complete emotional state due to potential ambiguity, contextual dependencies, or modality-specific noise [6]. For instance, facial expressions can be deliberately masked, text lacks paralinguistic information, and vocal tone can be misleading without semantic context, especially in real-world conditions involving noise or sarcasm. Integrating information from multiple modalities offers a promising path towards more comprehensive, robust, and accurate emotion recognition, mirroring human perceptual processes [6].

Motivated by the need for improved emotion recognition systems, this paper introduces the 'TERT-Ensemble' technique, focusing on tri-modal analysis using image, text, and audio data. Our methodology emphasizes a structured, two-stage approach. First, we focus on building highly reliable unimodal predictors. This involves systematically evaluating a diverse set of contemporary deep learning architectures—including Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), language Transformers (BERT, DeBERTa, RoBERTa), and sequence models like Long Short-Term Memory networks (LSTMs) [7]–[10]—on carefully aggregated benchmark datasets (CK+, FER-2013, RAF-DB; Twitter Emotions; CREMA-D, RAVDESS, SAVEE). To further enhance unimodal performance, we implement intra-modal ensembles, combining the predictions from the base models within each modality using weighted averaging based on their validation performance. This initial stage yielded robust unimodal ensemble predictors with weighted F1-scores of **77.98%** (Image), **73.69%** (Text), and **66.43%** (Audio) on their respective test sets.

The second stage involves "tri-modal late fusion". We combine the probabilistic outputs generated by the three optimized intra-modal ensembles using a weighted averaging strategy, where weights reflect the empirically determined reliability of each modality's ensemble predictor. This final fusion model was evaluated on a simulated tri-modal test set, constructed by pairing congruent emotional samples from the individual test sets due to the lack of a naturally aligned corpus matching our training data. On this simulated set, the tri-modal fusion achieved a high accuracy of **95.56%** and a weighted F1-score of **0.96**. An interactive demonstration interface using the best single unimodal models was also developed.

This work contributes through: (1) A systematic evaluation of architectures for unimodal emotion recognition. (2) Demonstration of intra-modal ensembling benefits. (3) Implementation and evaluation of a two-stage (intra-modal then tri-modal) late fusion model, showing high potential on simulated data. (4) Establishment of strong unimodal and fused performance baselines. This research supports UN SDGs 3 (Health), 9 (Innovation), and 4 (Education) [11]–[13]. The paper proceeds as follows: Section II reviews literature, Section III discusses key parameters, Section IV details the methodology, Section V describes datasets, Section VI covers the experimental setup, Section VII presents results, and Section VIII concludes.

II. Literature Survey and Limitations

Emotion recognition research has progressed significantly within individual modalities, although persistent challenges motivate multimodal approaches. In **FER**, deep learning models like CNNs and ViTs dominate over earlier handcrafted features [14], with ensembles enhancing robustness [15]. For **Text ER**, Transformers like BERT [9], RoBERTa [17], and DeBERTa [18] excel over RNNs/LSTMs [16], and ensembling them can offer further gains [19]. In **SER**, features like MFCCs [20] are classified using CNNs, LSTMs [21], or hybrid models [22], with ensembling often improving accuracy [23]. While **Multimodal ER** generally outperforms unimodal systems [6], [24], fusion is complex, ranging from early/late [25] to attention/transformer-based methods [26]. Late fusion remains a viable strategy when leveraging strong unimodal predictors [24]. Key limitations motivating our work include: **Data Constraints** (spontaneous data scarcity, acted data generalization limits) [22]–[24], [26], **Fusion Complexity** [24], [26], **Computational Cost** [15], [19], [23], [26], **Inherent Ambiguity** (sarcasm, speaker variability, masked expressions) [15], [19], [23], and **Inter-Class Similarity** [15], [22].

III. Parameters Affecting Performance

Model performance is contingent upon numerous factors: **1. Data Quality/Characteristics:** Dataset diversity, volume, class balance, annotation quality. **2. Preprocessing/Augmentation:** Image normalization/augmentation, text cleaning/tokenization, audio feature extraction/augmentation choices. **3. Model Architecture/Configuration:** Base model selection, pre-training, fine-tuning strategy, classifier head design. **4. Training Hyperparameters:** LR schedule, batch size, optimizer, epochs, early stopping, loss function, weighting strategies, regularization. **5. Ensemble/Fusion Configuration:** Both intra-modal (base model diversity, validation F1 weighting) and tri-modal (intra-modal ensemble test F1 weighting) strategies impact final results.

IV. Methodology: TERT-Ensemble Unimodal Analysis and Fusion

Our methodology adopted a two-stage fusion process, beginning with optimizing unimodal performance via intra-modal ensembling, followed by tri-modal late fusion. The workflow is illustrated in Fig. 1.

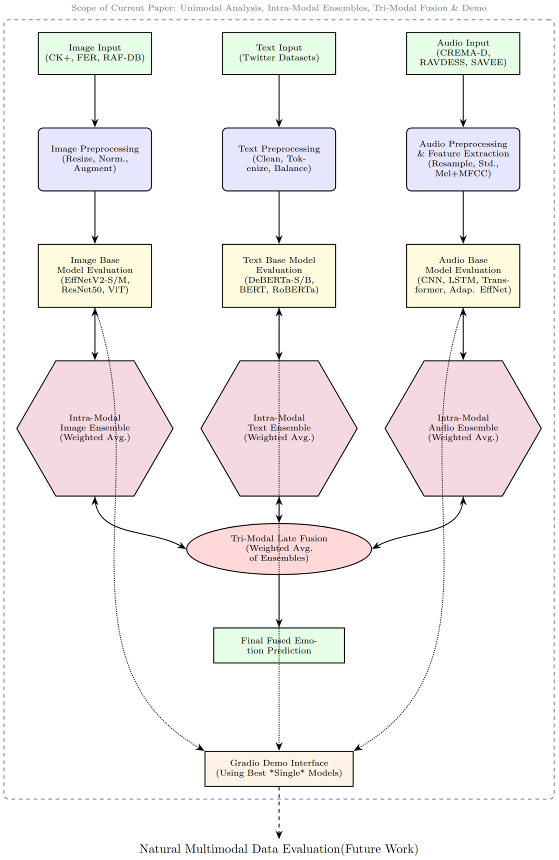

Scope of Current Paper: Unimodal Analysis, Intra-Modal Ensembles, Tri-Modal Fusion & Demo
Image Input (CK+, FER, RAF-DB)
Text Input (Twitter Datasets)
Audio Input (CREMA-D, RAVDESS, SAVEE)
Image Preprocessing (Resize, Norm., Augment)
Text Preprocessing (Clean, Tokenize, Balance)
Audio Preprocessing & Feature Extraction (Resample, Std., Mel+MFCC)
Image Base Model Evaluation (EffNetV2-S/M, ResNet50, ViT)
Text Base Model Evaluation (DeBERTa-S/B, BERT, RoBERTa)
Audio Base Model Evaluation (CNN, LSTM, Transformer, Adap. EffNet)
Intra-Modal Image Ensemble (Weighted Avg.)
Intra-Modal Text Ensemble (Weighted Avg.)
Intra-Modal Audio Ensemble (Weighted Avg.)
Tri-Modal Late Fusion (Weighted Avg. of Ensembles)
Final Fused Emotion Prediction
Gradio Demo Interface (Using Best *Single* Models)
Natural Multimodal Data Evaluation(Future Work)

Fig. 1. Methodology flowchart showing parallel unimodal evaluation leading to intra-modal ensembles, followed by tri-modal fusion and demo integration.

A. Unimodal Base Model Training

Multiple diverse architectures were independently trained for each modality using combined datasets and standardized procedures (detailed in Sec. VI). Model weights corresponding to the best validation weighted F1-score were saved for each architecture.

- **Image Models Trained:** EfficientNetV2-S/M [27], ResNet50 [28], ViT-B/16 [8].
- **Text Models Trained:** DeBERTa-v3-small/base [18], BERT-base [9], RoBERTa-base [17].
- **Audio Models Trained:** Custom 2D CNN, CNN-LSTM w/ Attention [30], CNN-Transformer, Adapted EfficientNetV2-S [27].

B. Intra-Modal Ensemble Fusion

For each modality, a fused prediction was generated by combining the outputs of its constituent trained models.

- **Technique:** Late Fusion via Weighted Averaging of Probabilities.
- **Process:** For a given input, softmax probabilities were obtained from all relevant trained models (e.g., all 4 image models for an image input).
- **Weighting:** Probabilities were averaged using weights (W_{intra}) derived from each base model's performance on the *validation* set (using weighted F1, normalized to sum to 1, see Sec VII-B logs for values).
- **Output:** A single probability vector ($P_{img\_ens}, P_{txt\_ens}, P_{aud\_ens}$) per modality.

These intra-modal ensembles were evaluated on their respective test sets (Results in Sec. VII-B).

C. Tri-Modal Late Fusion

The final fusion stage combined the outputs of the three optimized intra-modal ensembles.

- **Technique:** Late Fusion via Weighted Averaging of Intra-Modal Ensemble Probabilities.
- **Process:** For a tri-modal input (image, text, audio), the probability vectors $P_{img\_ens}, P_{txt\_ens}, P_{aud\_ens}$ were obtained from the respective intra-modal ensembles.
- **Weighting:** These vectors were combined using $P_{fused} = w_{img\_tri} P_{img\_ens} + w_{txt\_tri} P_{txt\_ens} + w_{aud\_tri} P_{aud\_ens}$. Tri-modal weights ($w_{tri}$) reflected the overall reliability of each intra-modal ensemble, based on their respective *test set* weighted F1-scores (ImageEns: 0.7798, TextEns: 0.7369, AudioEns: 0.6643), normalized ($w_{img\_tri} \approx 0.358, w_{txt\_tri} \approx 0.338, w_{aud\_tri} \approx 0.305$).
- **Prediction:** Final emotion based on $\arg\max(P_{fused})$.

D. Simulated Test Set for Tri-Modal Evaluation

Lacking an aligned natural tri-modal corpus, a simulated test set (1195 samples) was created by pairing same-label samples from the individual modality test sets, limited by the minimum count per class across modalities. This facilitates fusion evaluation but is a limitation.

E. Demonstration Interface

An interactive Gradio [51] interface was built using the best *single* unimodal models identified during the initial base model evaluation stage (EffNetV2-M, DeBERTa-v3-small, adapted EfficientNet).

V. Datasets Used

Publicly available datasets were combined per modality for 7 emotion classes.

A. Image Datasets

CK+ [31], FER-2013 [32], and RAF-DB [33] combined ($\tilde{5}$2k total samples). 'Contempt' mapped to 'disgust'. Sample diversity shown in Fig. 2.

B. Text Datasets

Emotion Detection from Text [34] and Emotions Dataset [35] combined. Labels mapped; class balancing and dummy data used ($\tilde{7}$3k samples).

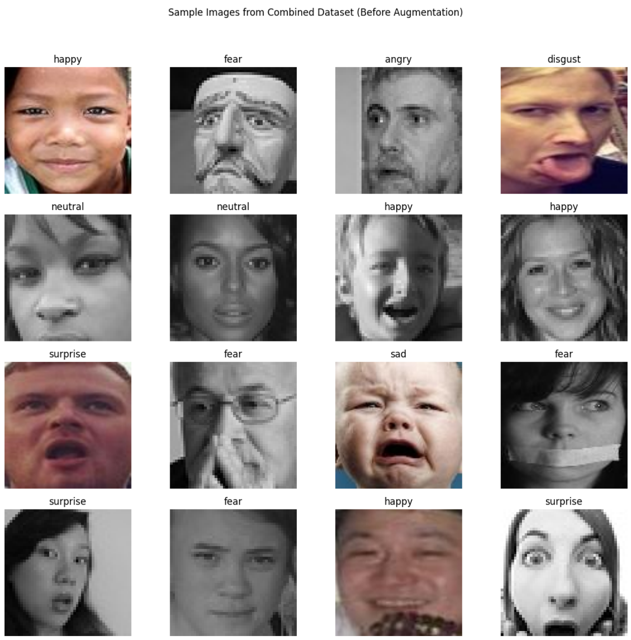


Fig. 2. Sample images from combined datasets.

C. Audio Datasets

CREMA-D [36], RAVDESS [37] (speech only, 'calm' to 'neutral'), and SAVEE [38] combined (9,362 files), mostly acted speech.

VI. Experimental Setup

A. Implementation Environment

Kaggle Notebooks, NVIDIA GPUs. Libraries: PyTorch [39], Transformers [40], torchaudio [41], librosa [42], pandas [43], NumPy [44], scikit-learn [45], Matplotlib [46], Seaborn [47], Albumentations [50], Gradio [51].

B. Training Procedures

- **Data Split:** Approx. 70/15/15 Train/Val/Test. Test Sizes: Image=7831, Text=10971, Audio=1405. Stratified.
- **Optimization:** AdamW [48] (WD: 1e-4 img/aud, 0.01 txt).
- **LR Scheduling:** OneCycleLR [49] (Max LR: Img 8e-4, Txt 5e-5, Aud 1e-3).
- **Loss Function:** Cross-Entropy. Class weights (img/aud); Label smoothing (0.1 img/txt).
- **Batching/Epochs:** Batch Sizes (Img 32, Txt 16, Aud 32). Grad Accum (Img 2, Txt 4, Aud 2). Max Epochs (Img 30, Txt 6, Aud 30).
- **Regularization:** Weight decay, Dropout, Gradient clipping (norm=1.0 txt/aud).
- **Mixed Precision:** 'torch.cuda.amp' used.
- **Early Stopping:** Based on validation weighted F1 (Patience: Img 10, Txt 3, Aud 10). Best model saved.

VII. Results, Metrics, and Discussion

A. Evaluation Metrics

Standard metrics were used: Accuracy, Precision, Recall, Weighted F1-score (primary), Macro F1-score, Confusion Matrices.

B. Intra-Modal Ensemble Performance

Ensembles combining multiple architectures within each modality were evaluated on their respective test sets.

1) Image Ensemble Results: The weighted average ensemble slightly outperformed the best single image model (EfficientNetV2-M), achieving a weighted F1 of 77.98% (Table I). The confusion matrix is shown in Fig. 3.

TABLE I
Image Modality: Best Single vs. Ensemble

Model	Test Acc.	Test F1 (Wgt)
Best Single (EffNetV2-M)	0.7621	0.7635
Intra-Modal Ensemble	**0.7794**	**0.7798**

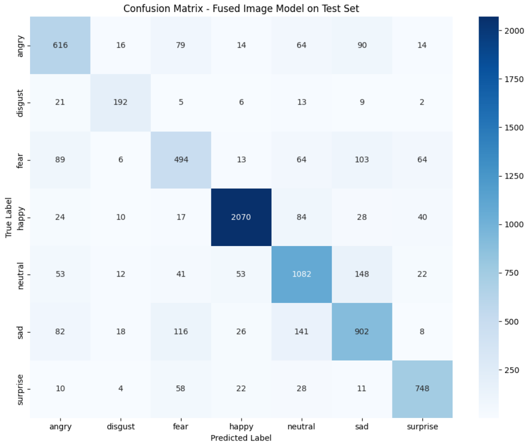


Fig. 3. Confusion Matrix (Intra-Modal Image Ensemble)

Discussion (Image Ensemble): The ensemble provided a notable improvement (+1.63% F1 absolute) over the best single model, indicating value in combining diverse vision architectures (CNNs and ViT) for FER.

2) Text Ensemble Results: The text ensemble also showed a modest improvement over the best single model (DeBERTa-v3-small), reaching 73.69% weighted F1 (Table II). Confusion matrix in Fig. 4.

TABLE II
Text Modality: Best Single vs. Ensemble

Model	Test Acc.	Test F1 (Wgt)
Best Single (DeBERTa-S)	0.7249	0.7292
Intra-Modal Ensemble	**0.7302**	**0.7369**

Discussion (Text Ensemble): The gain (+0.77% F1 absolute) was smaller than for images, possibly due to the relative similarity of the base Transformer models.

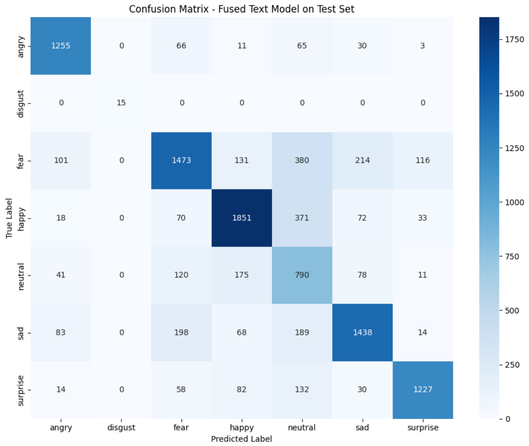


Fig. 4. Confusion Matrix (Intra-Modal Text Ensemble)

3) Audio Ensemble Results: The audio ensemble provided only marginal improvement over the best single model (LSTM), achieving 66.43% weighted F1 (Table III). Confusion matrix in Fig. 5.

TABLE III
Audio Modality: Best Single vs. Ensemble

Model	Test Acc.	Test F1 (Wgt)
Best Single (LSTM)	0.6676	0.6634
Intra-Modal Ensemble	**0.6698**	**0.6643**

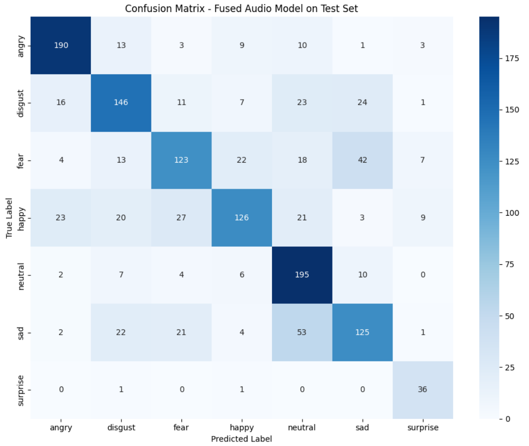


Fig. 5. Confusion Matrix (Intra-Modal Audio Ensemble)

Discussion (Audio Ensemble): The minimal improvement suggests the inclusion of the poorly performing Transformer likely limited the ensemble's effectiveness.

C. Tri-Modal Fusion Performance

The final fusion model combined the outputs of the three intra-modal ensembles using weights derived from their test F1 scores. Performance on the simulated test set is in Table IV.

TABLE IV
TRI-MODAL FUSION MODEL PERFORMANCE (SIMULATED TEST SET)

Metric	Value
Accuracy	95.56%
Weighted F1-Score	0.96
Macro F1-Score	0.96

Figs. 6 and 7 show the confusion matrix and ROC curves.

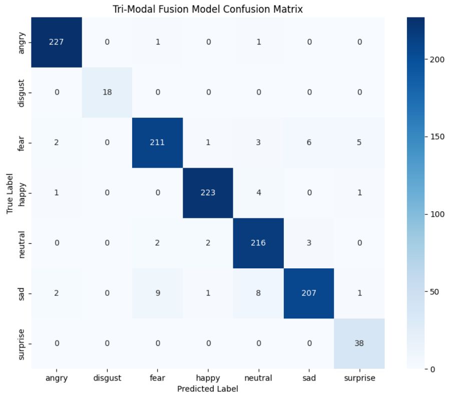


Fig. 6. Confusion Matrix (Tri-Modal Fusion Model - Simulated Test Set)

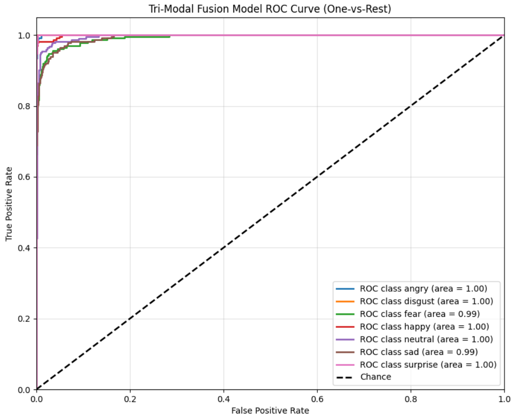


Fig. 7. ROC Curves (Tri-Modal Fusion Model - Simulated Test Set)

Discussion (Tri-Modal Fusion): Exceptionally high performance (95.56% Acc, 0.96 F1) was achieved. This strongly validates the fusion approach's potential but must be interpreted cautiously due to the simulated test set containing congruent emotional cues across modalities, likely simplifying the task compared to real-world data.

D. Sample Predictions Visualization

Qualitative results comparing intra-modal ensemble predictions and the final tri-modal fusion output on sample data are illustrated in Figs. 8-10.

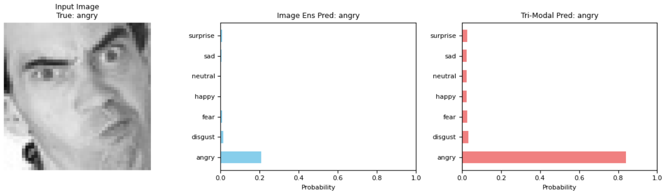


Fig. 8. Sample Image Predictions (Intra-Modal Ensemble vs. Tri-Modal Fused).

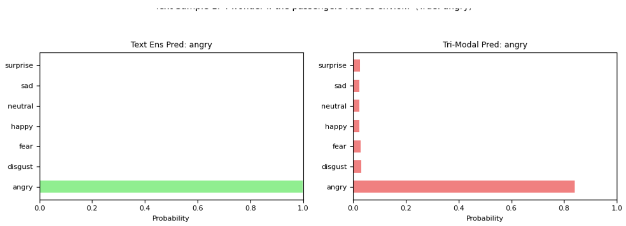


Fig. 9. Sample Text Predictions (Intra-Modal Ensemble vs. Tri-Modal Fused).

E. Demonstration Interface

An interactive Gradio interface [51] (Figs. 11-13) using the best *single* unimodal models provides a practical demonstration of individual component capabilities.

F. Overall Discussion

This study successfully developed robust unimodal predictors via intra-modal ensembling (Image F1: 77.98%, Text F1: 73.69%, Audio F1: 66.43%), generally improving over the best single models. A subsequent tri-modal late fusion model combining these ensembles achieved high accuracy (95.56%) and F1 (0.96) on simulated congruent data, confirming the potential of the TERT-Ensemble fusion approach. However, the reliance on simulated data requires cautious interpretation and

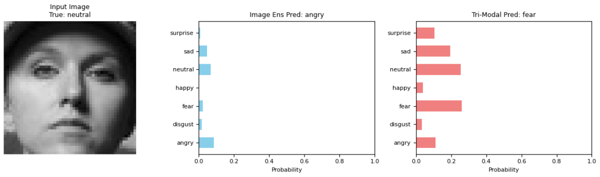
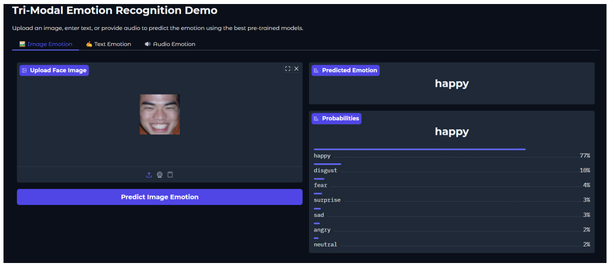


Fig. 10. Sample Audio Predictions (Intra-Modal Ensemble vs. Tri-Modal Fused).


Tri-Modal Emotion Recognition Demo

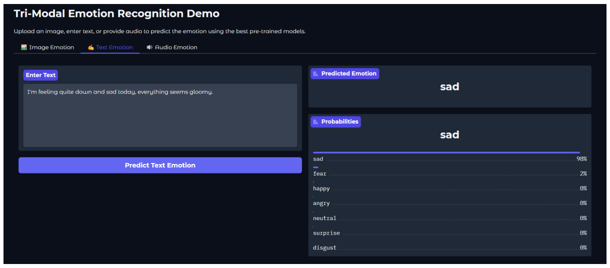
Upload an image, enter text, or provide audio to predict the emotion using the best pre-trained models.

happy

happy

```
happy       77%
disgust     10%
fear         4%
surprise     3%
sad          3%
angry        2%
neutral      2%
```

Fig. 11. Gradio Demo Interface: Image Emotion Tab.



Tri-Modal Emotion Recognition Demo

Upload an image, enter text, or provide audio to predict the emotion using the best pre-trained models.

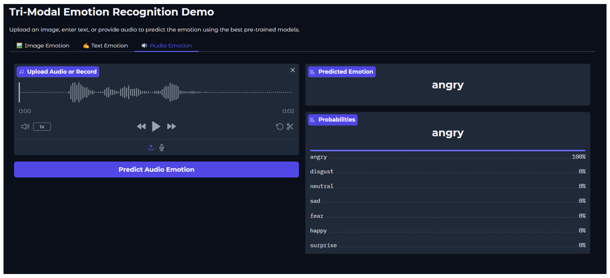
I'm feeling quite down and sad today, everything seems gloomy.

Predicted Emotion: sad

Probabilities: sad

```
sad ........ 98%
fear ........ 2%
happy ....... 0%
angry ....... 0%
neutral ..... 0%
surprise .... 0%
disgust ..... 0%
```

Fig. 12. Gradio Demo Interface: Text Emotion Tab.



Tri-Modal Emotion Recognition Demo

Upload an image, enter text, or provide audio to predict the emotion using the best pre-trained models.

Predicted Emotion: angry

Probabilities: angry

```
angry ..... 100%
disgust ..... 0%
neutral ..... 0%
sad ......... 0%
fear ........ 0%
happy ....... 0%
surprise .... 0%
```

Fig. 13. Gradio Demo Interface: Audio Emotion Tab.

necessitates future validation on natural multimodal datasets. The Gradio demo provides a functional showcase.

VIII. Conclusion and Future Work

This paper presented the TERT-Ensemble approach, detailing the development of optimized unimodal emotion predictors through intra-modal ensembling and their subsequent combination via tri-modal late fusion. We established strong unimodal ensemble baselines (Image F1: 77.98%, Text F1: 73.69%, Audio F1: 66.43%) and demonstrated the high potential of the final fusion stage (95.56% accuracy, 0.96 F1) on simulated data. An interactive demo illustrates component functionality. This work provides a validated methodology and robust benchmarks. Future research will focus on evaluating the tri-modal fusion model on naturalistic multimodal datasets and exploring advanced fusion techniques to better handle potential cross-modal conflicts and dependencies.

References

[1] R. W. Picard, *Affective computing*. Cambridge, MA, USA: MIT Press, 1997.

[2] A. Mehrabian, *Nonverbal communication*. London, UK: Routledge, 2017.

[3] Z. Li and S. Li, "Deep Facial Expression Recognition: A Survey," *arXiv preprint arXiv:1804.08348*, 2018.

[4] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, Sep. 2017.

[5] B. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, May 2018.

[6] G. Ramakrishnan and S. K. Saha, "A Survey on Multimodal Emotion Recognition using Deep Learning," *arXiv preprint arXiv:2103.03139*, 2021.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[8] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL-HLT)*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186. [Online]. Available: https://arxiv.org/abs/1810.04805

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[11] G. Muhammad, M. Alhamid, M. Alsulaiman, and A. A. Al-Rodhaan, "EEG Signal Analysis for Detecting Depression Using Machine Learning Techniques: A Review," *IEEE Access*, vol. 9, pp. 114569–114580, 2021,

[12] A. Ochuba et al., "Leveraging Artificial Intelligence to Meet the Sustainable Development Goals," *ResearchGate*, Jan. 2024. [Online].

[13] S. K. D'Mello and A. Graesser, "Affect detection from spoken language," *Speech Communication*, vol. 53, no. 9-10, pp. 1160–1174, Nov.–Dec. 2011,

[14] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, Jan. 1996.

[15] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A Deep Multi-Task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.

[16] E. Cambria, "Affective Computing and Sentiment Analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.

[17] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.

[18] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2021. [Online]. Available: https://arxiv.org/abs/2006.03654

[19] F. A. Acheampong, W. Nunoo-Mensah, and A. Aggrey, "Transformer-Based Ensemble Model for Sentiment Analysis of COVID-19 Tweets," *Applied Sciences*, vol. 11, no. 15, Art. no. 6752, Jul. 2021.

[20] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the Munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia (MM '10)*, Florence, Italy, Oct. 2010, pp. 1459–1462.

[21] M. Z. P. Ishaq, S. M. S. Ahmad, H. S. Munawar, and A. S. Abdullah, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," *IEEE Access*, vol. 9, pp. 109033–109056, 2021.

[22] G. Trigeorgis et al., "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5200–5204.

[23] X. Li, D. Tu, L. Zhang, X. Wang, and B. Xu, "SER-CNNs: An Ensemble of CNNs for Speech Emotion Recognition," *IEEE Access*, vol. 8, pp. 101196–101205, 2020.

[24] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, Sep. 2017.

[25] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. 13th ACM Int. Conf. Multimedia (MM '05)*, Singapore, Nov. 2005, pp. 399–402.

[26] Y.-H. H. Tsai et al., "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Florence, Italy, Jul. 2019, pp. 6558–6569.

[27] M. Tan and Q. V. Le, "Efficientnetv2: Smaller models and faster training," in *Proc. Int. Conf. Machine Learning (ICML)*, PMLR, vol. 139, Jul. 2021, pp. 10096–10106.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[29] D. S. Park et al., "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 2613–2617.

[30] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015.

[31] P. Lucey et al., “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression,” in *Proc. 2010 IEEE Computer Soc. Conf. Computer Vision Pattern Recognition-Workshops (CVPRW)*, San Francisco, CA, USA, Jun. 2010, pp. 94–101.

[32] I. J. Goodfellow et al., “Challenges in representation learning: A report on the 2013 ICML workshop,” *arXiv preprint arXiv:1312.6082*, 2013.

[33] S. Li, W. Deng, and J. Du, “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,” in *Proc. IEEE Conf. Computer Vision Pattern Recognition (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2252–2260.

[34] J. D. Perez, “Emotion Detection from Text,” Kaggle Dataset, 2020. [Online]. Available: https://www.kaggle.com/datasets/praveengovi/emotions-dataset-for-nlp

[35] S. Saravia, H. C. C. Liu, Y. Huang, J. Wu, and Y. A. Chen, “CARER: Contextualized Affect Representations for Emotion Recognition,” in *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*, Brussels, Belgium, Nov. 2018, pp. 3687–3697.

[36] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, “Crema-d: Crowd-sourced emotional multimodal actors dataset,” *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Oct.–Dec. 2014.

[37] S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions,” *PLoS ONE*, vol. 13, no. 5, p. e0196391, May 2018.

[38] S. Haq and P. J. B. Jackson, “Speech emotion recognition using spectrogram phoneme embedding,” in *Proc. 2009 Ninth IEEE Int. Conf. Data Mining Workshops (ICDMW ’09)*, Miami, FL, USA, Dec. 2009, pp. 41–46.

[39] A. Paszke et al., “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, 2019, vol. 32.

[40] T. Wolf et al., “Transformers: State-of-the-art natural language processing,” in *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP): System Demonstrations*, Online, Oct. 2020, pp. 38–45.

[41] H. Yang et al., “Torchaudio: An audio library for PyTorch,” in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 7482–7486.

[42] B. McFee et al., “librosa: Audio and music signal analysis in python,” in *Proc. 14th Python Sci. Conf.*, Austin, TX, USA, 2015, vol. 8, pp. 18–25.

[43] W. McKinney, “Data Structures for Statistical Computing in Python,” in *Proc. 9th Python Sci. Conf.*, Austin, TX, USA, 2010, pp. 56–61.

[44] C. R. Harris et al., “Array programming with NumPy,” *Nature*, vol. 585, pp. 357–362, Sep. 2020.

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

[46] J. D. Hunter, “Matplotlib: A 2D graphics environment,” *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, May/Jun. 2007.

[47] M. L. Waskom, “Seaborn: statistical data visualization,” *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, Apr. 2021.

[48] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, New Orleans, LA, USA, May 2019.

[49] L. N. Smith, “A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay,” *arXiv preprint arXiv:1803.09820*, 2018.

[50] A. Buslaev et al., “Albumentations: fast and flexible image augmentations,” *Information*, vol. 11, no. 2, p. 125, Feb. 2020.

[51] A. Abid et al., “Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild,” *arXiv preprint arXiv:1906.02569*, 2019.