Lightweight Multimodal Emotion Recognition Using Cross-Dataset Feature Fusion of Text and Facial Expressions

Susrita Mishra
Dept. of Computer Science & Engineering
NIST University
Berhampur, India
susrita.mishra.cse.2022@nist.edu

Phalguni Patnaik
Dept. of Computer Science & Engineering
NIST University
Berhampur, India
phalguni.patnaik.cse.2022@nist.edu

Bandhan Panda
Dept. of Computer Science & Engineering
NIST University
Berhampur, India
bandhan.panda@nist.edu

Samikhya Patnaik
Dept. of Computer Science & Engineering
NIST University
Berhampur, India
samikhya.patnaik.cse.2022@nist.edu

Ujjwal Singh
Dept. of Computer Science & Engineering
NIST University
Berhampur, India
ujjwal.singh.cse.2022@nist.edu

Santosh Kumar Kar
Dept. of Computer Science & Engineering
NIST University
Berhampur, India
santoshkumarkar@nist.edu

How to Cite this Article:

Mishra, S., Patnaik, P., Patnaik, S., Singh, U., Kar, S. K. & Panda, B. (2026). Lightweight Multimodal Emotion Recognition Using Cross-Dataset Feature Fusion of Text and Facial Expressions. International Journal of Creative and Open Research in Engineering and Management, 02(03). https://doi.org/10.55041/ijcope.v2i3.142

https://doi.org/10.55041/ijcope.v2i3.142

Abstract- Emotion recognition has become a significant undertaking in affective computing that makes intelligent systems recognise and act upon human emotions. It is well involved in human-computer interaction, healthcare monitoring, and smart virtual assistants. Most conventional emotion recognition systems currently are based on a single expressive system, e.g., textual affect or facial expression, and can only be used to a limited degree to understand the depth and multi-proponent characteristics of human feelings. To address this shortcoming, this paper proposes a multimodal emotion recognition framework that integrates text and visual cues at the deep feature level. The proposed system utilises DistilBERT to extract textual representations of the context of the ISEAR dataset and EfficientNet-B3 to extract facial expression features of the FER2013 and RAF-DB datasets. Because the textual and visual data are sampled across datasets, a cross-dataset pairing strategy is proposed to form multimodal training samples by matching textual descriptions to facial images with the matching emotion labels. The obtained features are fused through a feature fusion mechanism that is a gated one and fed into a Long Short-Term Memory (LSTM) classifier. Experimental findings indicate that the proposed multimodal model has an accuracy of 82%, performing better than the text-only model (61%) and the image-only model (69%), which proves the applicability of multimodal emotion recognition that was cross-dataset trained.

Keywords- *Multimodal Emotion Recognition, Affective Computing, DistilBERT, EfficientNet, Feature Fusion, Deep Learning.*

I. INTRODUCTION

Another important field of research in affective computing involves emotion recognition, which is the process of enabling intelligent systems to perceive and react to the emotions of human beings. It has applications in human-computer interaction, health care monitoring, virtual assistants and smart learning systems. Thus, single-mode recognition of emotion has a high probability of missing out on the other emotional contexts. Recent research indicates that multimodal emotion recognition systems involving a combination of multiple modalities are capable of significantly enhancing performance relative to traditional unimodal methods of emotion recognition [1], [2]. The experimental developments in natural language processing and computer vision have made it possible to develop better emotion recognition systems. BERT and its smaller variant DistilBERT are models based on transformers that can transform textual emotion recognition by means of strong contextual representations [5], [6].

Nevertheless, many of the current multimodal emotion recognition systems are based on synchronised multimodal data sets, which can be difficult and costly to obtain. To overcome this challenge, this paper develops a lightweight multimodal emotion recognition system that combines text modality and visual modality emotional data using independent data sets. The suggested system applies DistilBERT to reconstruct the textual characteristic of the ISEAR dataset, and EfficientNet-B3 reconstructs the facial emotion characteristic of the FER2013 and RAF-DB datasets.

To create multimodal samples, a cross-dataset pairing approach is used, and the features are combined via a gated fusion system and then through an emotion-predicting LSTM classifier. The key contributions of this work are the construction of the cross-dataset multimodal pairing strategy that correlates text emotion descriptions provided in the ISEAR dataset with facial images of the FER2013 dataset and RAF-DB dataset, and makes multimodal learning possible without synchronised data. Besides that, a lightweight multimodal architecture is offered by combining DistilBERT to extract textual features and EfficientNet-B3 to extract visual features with the frozen pretrained encoders to save on computational complexity. It is also the framework that introduces a gated feature fusion mechanism to sequentially blend textual and visual embeddings and learn opposing emotional representations. Lastly, the feasibility of the proposed method is tested on three emotion databases and shows better results than unimodal emotion recognition frameworks.

II. RELATED WORK

Recent studies of emotion recognition have targeted multimodal strategies of combining various sources of emotional information. Wafa et al. presented a multimodal emotion recognition system integrating text, audio, video, and motion information with complex deep learning methods to enhance emotion recognition accuracy [1]. Likewise, El Maazouzi and Retbi designed a multimodal system to combine textual and visual information to identify emotional and cognitive states in the e-learning settings [2]. Research conducted in the field of affective computing indicates that multimodal systems are more reliable in the recognition of emotions than unimodal system, by utilising complementary cues of emotions [3].

The progress in natural language processing has made a great breakthrough in detecting emotions through text. BERT proposed bi-directional transformer architectures to understand language in contexts [5], whereas DistilBERT proposes a sparse variant of BERT that simplifies computational operations but does not decrease the performance [6]. EfficientNet is an effective image recognition technique that uses optimised network scaling schemes to enhance image recognition [7] in the field of computer vision, and deep learning techniques have been shown to be effective in facial expression recognition tasks [8]. Facial emotion recognition in the real world has also been demonstrated by large-scale datasets like AffectNet [9], and prior studies demonstrated the applicability of deep neural networks to learn the representation of facial emotions [10].

There are a number of multimodal fusion methods that have been suggested to combine emotional cues across various modalities. Multi-modal sentiment analysis networks are based on interactions between two or more modalities to analyze sentiment [11], and multimodal sentiment analysis research emphasizes the relevance of integrating visual and textual information to understand emotions better [12]. IEMOCAP and other multimodal datasets offer useful data in

understanding emotional interactions [13]. The Long Short-Term Memory networks used as sequential models remember representations of emotional cues across a period of time [14], whereas the transformer architectures additionally improve representation learning with attention mechanisms [15]. A large number of optimisation algorithms like Adam are popular in training deep neural networks [16], and sentiment analysis studies offer practical methods of finding emotional information in text-based inputs [17].

III. RESEARCH GAP AND MOTIVATION

The recent studies of emotion recognition have demonstrated that multimodal techniques can be used to enhance accuracy and strength of emotion detection systems by combining various emotional signal sources, including text, images, and audio [1], [2]. Affective computing research has suggested that using complementary modalities will enable systems to express more detailed emotional representations than when using unimodal methods [3]. In the same way, multimodal machine learning studies have advocated the significance of representation learning and fusion methods in the combination of heterogeneous data sources [4]. In spite of these developments, a large number of current multimodal emotion recognition systems utilise complex architectures and demand large synchronised multimodal datasets, which are challenging to collect and label.NLP innovations have produced effective transformer-based models like BERT and DistilBERT that can readily access contextual emotion data via text [5], [6]. Simultaneously, deep convolutional neural networks like EfficientNet have shown high effectiveness in the recognition of facial emotions using visual features that are discriminative to face images [7]. Moreover, single-modality emotion recognition models based on large-scale datasets of facial expressions have been developed and can be used in the real-world setting [8]. Nevertheless, most of the current research is either single-modality or is based on entirely aligned multimodal datasets. This constrains the scalability and feasibility of multimodal systems to real-world conditions where synchronised multimodal data is not necessarily present.

Moreover, employing a single dataset on facial expressions can potentially diminish the variability of visual emotion representation. Because of these issues, this research suggests a lightweight multimodal emotion recognition system that combines textual and visual emotional expression by using separately gathered datasets. The suggested method involves the combination of the textual descriptions of emotions in the ISEAR dataset with the facial images in FER-2013 and RAF-DB through the cross-dataset pairing approach. Using transformer-based text representations and deep convolutional visual features with gated feature fusion, the framework will be useful in enhancing emotion recognition performance without compromising computational efficiency. The main originality of the given work is the use of cross-dataset multimodal fusion that makes it possible to learn multimedia without having mutually synchronized multimodal datasets.

IV. METHODOLOGY

The proposed framework performs multimodal emotion recognition by integrating textual and visual emotional cues using deep learning models, as shown in Fig.1.

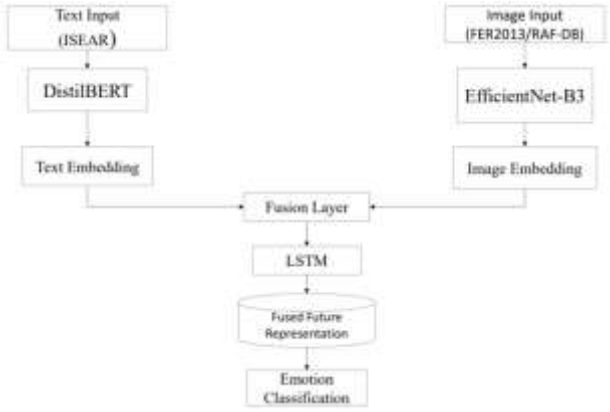

Fig. 1. Architecture of the Proposed Multimodal Emotion

The system combines pretrained encoders for text and facial images, followed by a multimodal fusion mechanism and a classification network to predict the final emotion category. The overall pipeline consists of five main stages: data preprocessing and multimodal pairing, text feature extraction, visual feature extraction, multimodal feature fusion, and emotion classification.

1) Data Preprocessing and Multimodal Pairing

Three datasets are used in this work- ISEAR dataset for textual emotion descriptions, the FER2013 dataset for facial expression images, RAF-DB dataset for facial emotion recognition.

The ISEAR dataset contains text descriptions labelled with emotion categories such as joy, anger, sadness, fear, and disgust. The FER2013 and RAF-DB datasets contain facial images annotated with basic emotion classes. Since the textual and visual datasets are independently collected, a cross-dataset pairing strategy is applied. For each text sample in the ISEAR dataset, a facial image from the FER2013 dataset with the same emotion label is selected to form a multimodal pair. The RAF-DB dataset is used to improve the robustness of facial emotion representation during visual feature learning. Before training, textual data is tokenised and encoded, while facial images are resized and normalised to ensure consistent input representation.

2) *Text Feature Extraction*

Textual emotion features are extracted using DistilBERT, a lightweight transformer-based language model. The input text sequence is tokenised and converted into contextual embeddings using the DistilBERT tokeniser. These embeddings capture semantic and emotional information from the textual description. The final hidden representation of DistilBERT is used as the textual feature vector, which serves as the input to the multimodal fusion layer.

3) *Visual Feature Extraction*

Facial emotion features are extracted using an EfficientNet-B3 convolutional neural network. The model processes facial images to learn discriminative visual patterns related to emotional expressions such as mouth shape, eye movement, and eyebrow position. Images from both FER2013 and RAF-DB datasets are used to obtain robust facial emotion representations. The EfficientNet encoder outputs a high-dimensional visual embedding representing the emotional characteristics of the facial image.

4) Multimodal Feature Fusion

To combine textual and visual information, a gated feature fusion mechanism is employed. Let H_t denote the textual embedding and H_v denote the visual embedding.

The multimodal representation is computed as:

$$F = g \odot H_v + (1 - g) \odot H_t$$

where g is a learnable gating vector and $\odot$ represents element-wise multiplication. The gating mechanism dynamically adjusts the contribution of textual and visual modalities during fusion, allowing the model to learn complementary emotional representations.

5) *Emotion Classification*

The fused multimodal representation is passed to a **L**ong Short-Term Memory (LSTM) network, which models contextual dependencies in the combined feature space. Although the input corresponds to a single timestep, the LSTM layer helps capture relationships within the fused representation. The output of the LSTM is then passed through a fully connected layer followed by a softmax classifier to predict the final emotion category.

The LSTM layer can also model non-linear interaction between textual and visual embeddings, in that the network is able to formulate the complex dependencies between features in the fused multimodal representation.

6) Training Strategy

To reduce computational complexity and preserve pretrained feature representations, the DistilBERT and EfficientNet encoders are frozen during training. Only the multimodal fusion layer, LSTM network, and final classification layer are updated during optimisation. The model is trained using the cross-entropy loss function and optimised using the Adam optimiser. This strategy enables efficient training while maintaining strong feature extraction capabilities.

A. Model Formulation

Let the input text sequence be represented as

$\mathbf{X} = \{w_1, w_2, \ldots, w_n\}$

1) where w_i denotes the i^{th} word token. The textual features are extracted using the DistilBERT encoder:

$H_t = f_t(\mathbf{X})$

where H_t represents the textual embedding.

2) For the visual modality, let the input facial image be $I \in R^{H \times W \times C}$

3) Visual features are extracted using EfficientNet:

$H_v = f_v(I)$

4) Gated Fusion:

$$F = g \odot H_v + (1 - g) \odot H_t$$

5) The fused representation is passed through an LSTM network:

$h_t = \text{LSTM}(F_t, h_{t-1})$

6) The final emotion prediction is generated using a softmax classifier:

y = Softmax (Wh + b)

where y represents the predicted emotion class.

C. Proposed Algorithm

Algorithm 1: Multimodal Emotion Recognition Using Cross-Dataset Feature Fusion

Input: Text dataset T = $\{t_i\}$(ISEAR), Image datasets I = $\{i_j\}$(FER2013, RAF-DB), Emotion labels Y, Text encoder $f_t(\cdot)$(DistilBERT), Visual encoder $f_v(\cdot)$ (EfficientNet-B3), Learning rate η, Number of epochs E

Output: Trained multimodal model parameters W*

1. Load the textual dataset T and facial image datasets I
2. Preprocess datasets
   2.1. Tokenise and encode text samples using DistilBERT tokeniser
   2.2. Resize and normalise facial images
3. Initialise model components
   3.1. Load pretrained DistilBERT encoder $f_t(\cdot)$
   3.2. Load pretrained EfficientNet-B3 encoder $f_v(\cdot)$
   3.3. Initialise gated fusion layer
   3.4. Initialise LSTM classifier
4. Freeze parameters of DistilBERT and EfficientNet encoders
5. Create multimodal training pairs for each text sample t_i in dataset T: select facial image i_j from FER2013 with the same emotion label
6. for epoch = 1 to E do
7. for each minibatch (t_i, i_j, y) do
8. $H_t = f_t(t_i)$
9. $H_v = f_v(i_j)$
10. $F = g \odot H_v + (1 - g) \odot H_t$
11. $h = LSTM(F)$
12. $\hat{y} = \widehat{Softmax}(Wh + b)$
13. $L = CrossEntropy(\hat{y}, y)$
14. $W \leftarrow W - \eta \nabla_W L$
15. end for
16. end for
17. Return trained parameters W*

TABLE I. NOTATIONS AND DESCRIPTIONS

Notation	Description
$T = \{t_i\}$	Text dataset (ISEAR) containing emotion descriptions
$I = \{i_j\}$	Image datasets (FER2013, RAF-DB) containing facial expressions
t_i	Input text sample
i_j	Input facial image
Y	Emotion label set
$f_t(\cdot)$	Text encoder (DistilBERT)
$f_v(\cdot)$	Visual encoder (EfficientNet-B3)
H_t	Text feature embedding
H_v	Visual feature embedding
F	Fused multimodal feature
g	Gating weight in feature fusion
$\odot$	Element-wise multiplication
h	LSTM hidden representation
W, b	Classifier parameters
$\hat{y}$	Predicted emotion label
y	True emotion label
L	Cross-entropy loss
η	Learning rate
E	Number of training epochs
W*	Final trained model parameters

Table I summarizes the symbols and variables used in the proposed mathematical formulation and Algorithm 1.

V. DATASET AND EXPERIMENTAL SETUP

Three datasets available in the public were taken to assess the proposed multimodal emotion recognition framework by repairing both textual and visual emotional data. ISEAR Dataset: The ISEAR dataset has textual reports on emotional experiences coded on the basis of the emotions of joy, anger, sadness, fear, disgust, guilt, and shame. DistilBERT tokeniser is used to tokenise the text samples and run them with the contextual embedding purpose. FER-2013 Dataset: FER-2013 dataset contains grayscale facial images that are tagged with emotion labels, such as happiness, sadness, anger, fear, surprise, disgust and neutral. The visual images are resized and normalized, then used in extracting the visual features. RAF-DB Dataset: Real-world Affective Faces Database (RAF-DB) consists of a wide variety of images of faces that are annotated by basic emotion classes, and is applied to strengthen the performance of facial emotion recognition. As the sets of text and images are not obtained together, a cross-

dataset pairing method is used where the ISEAR text samples are paired with FER-2013 images with the same emotion tag, whereas RAF-DB enhances the learning of visual features.

The implementation of the model is via the PyTorch language, with DistilBERT learning the textual features and EfficientNet-B3 learning the visual ones. The training process is followed by freezing the pretrained encoders, and the only layers optimised are the multimodal fusion layer, the LSTM classifier, and the final classification layer. The model is trained with an optimiser of Adam using a 0.0001 learning rate and 32 as the batch size, and is evaluated on the basis of accuracy, precision, recall, and F1-score.

The data sets that were utilised in this research give both the textual and the visual expression of emotions. The ISEAR dataset consists of about 7,666 textual emotion samples, both of which are split to include 20% as a test and 80% as a training dataset. FER2013 has 35,887 images, of which 80% are used as training and 20% as testing, respectively. RAF-DB has a total number of 29,672 facial images, of which an approximate of 23,737 is used to train, and 5,935 is utilised to test. Those datasets offer a variety of emotional representations to be used in the training and testing of the suggested multimodal-based emotion recognition.

VI. RESULTS AND DISCUSSION

This section evaluates the performance of the proposed multimodal emotion recognition framework that integrates textual and visual emotional information. The experiments analyse the effectiveness of multimodal learning compared with unimodal approaches. Three model configurations are evaluated: a text-only model using DistilBERT, an image-only model using EfficientNet-B3, and the proposed multimodal model combining both modalities through gated feature fusion and an LSTM classifier.

TABLE II. MODEL PERFORMANCE COMPARISON

Model	Dataset	Accuracy	Precision	Recall	F1-score
Text-only Model	ISEAR	0.61	0.60	0.59	0.60
Image-only Model	FER2013 +RAF-DB	0.69	0.68	0.67	0.69
Proposed Multimodal Model	ISEAR + FER2013 +RAF-DB	0.82	0.81	0.82	0.81

Table II shows the quantitative analysis of the multimodal emotion recognition model, image-only, and text-only models. The multimodal framework has the highest accuracy of 82, which is better than the text-only (61) and image-only (69) models. This advancement shows that the integration of both the textual and visual emotional indicators can be used to get the model to extract complementary emotional information and enhance the performance of such a model in terms of classification.

A. Text-Only Model Performance

As shown in Fig. 2, the initial experiment is an emotion recognition task based on text alone, using the ISEAR dataset. A contextual text embedding is obtained using DistilBERT, and then a prediction layer is used to predict emotions by classifying them. The confusion matrix displays the distribution of the predicted and real emotion classes.

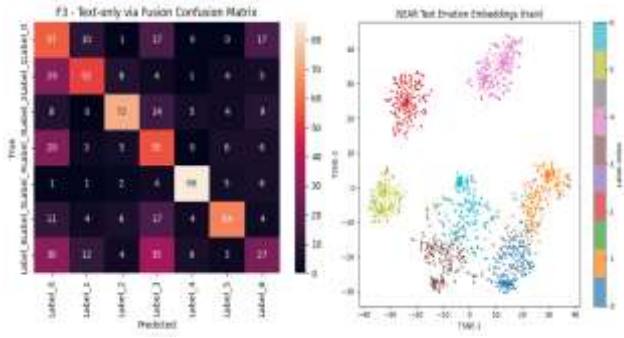

Fig. 2. Visualisation of emotion classification results using a confusion matrix and t-SNE embedding plot, illustrating DistilBERT's ability to separate different emotion classes in the ISEAR dataset.

Textual descriptions give effective emotional mentions that allow the model to distinguish several emotions, but according to the confusion matrix, some of the emotions that are similar in terms of facial expression, like fear and surprise, are sometimes misclassified.

B. Image-Only Model Performance

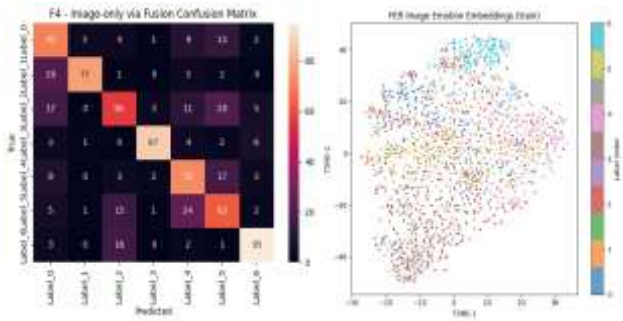

Fig. 3. Visualisation of facial emotion recognition results using t-SNE image embeddings and a confusion matrix, demonstrating EfficientNet's capability to distinguish different emotion classes from FER images and RAF-DB datasets.

The second study assesses the recognition of facial emotions based on visual data of the FER2013 and the RAF-DB collections. The EfficientNet-B3 convolutional neural network is used to extract the visual emotion features with the help of the facial images. The confusion matrix indicates that the model has the capabilities to identify facial expressions related to various emotions. The model works well on emotions that have a specific facial pattern, like happiness and surprise; some emotions of similar appearance (visually) may at times be misunderstood. Multimodal model, however, minimises this confusion by adding textual descriptions of contextual information, as demonstrated in Fig. 3.

C. Multimodal Model Performance

The final experiment will be a test of the suggested multimodal emotion recognition framework. Text and visual features that have been obtained with the help of DistilBERT and EfficientNet-B3, respectively, are fed into a gated feature fusion scheme, and the obtained fused representation is subjected to an LSTM classifier. The confusion matrix demonstrates that classification performance with respect to unimodal models is better.

The system is able to record complementary emotional clues by integrating both contextual textual and facial responses, which results in more reliable predictions, as in Fig. 4.

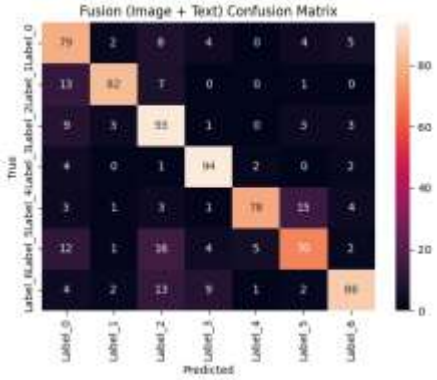

Fig. 4. Confusion matrix of the proposed multimodal model combining textual and visual features through gated fusion.

D. Multimodal vs Unimodal Comparison

In order to study the advantages of multimodal learning, the comparison of the results of the three models is conducted by means of a bar chart. The multimodal model is more accurate than the text-only and images-only ones. Although contextual emotion is only captured using language in the text-only model and facial expression to make decisions in the image-only model, both unimodal models' future capture only part of the emotive information. The multimodal model will combine the two modalities so that more emotional values can be represented.

TABLE III. ABLATION RESULTS

Model Configuration	Description	Accuracy
Text-only	DistilBERT using the ISEAR dataset	0.61
Image-only	EfficientNet-B3 using FER2013 and RAF-DB	0.69
Fusion (No Gating)	Concatenation fusion	0.77
Fusion (Gated)	Proposed model	0.82

Table III shows that the ablation experiment assesses the role of each modality in the performance. The text-only model is used to get contextual emotional information using text description, and the image-only model retrieves facial expression features. A multimodal model can perform effectively because the two modalities are combined using gated feature fusion.

The findings indicate that the combination of textual and visual features increases emotion recognition performance. DistilBert represents the textual meaning of emotion based on its context, whereas EfficientNet-B3 represents the features of facial expressions in photographs. The fusion mechanism of feature gating equalises the input of both modalities, and the LSTM classifier learns relationships in the fused representation.

This is enhanced by multimodal fusion which gives contextual emotional information during textual description and visual indications of facial images. Emotions like fear and surprise that can hardly be differentiated via the facial features alone are useful in multimodal fusion. Nonetheless, the pairing of cross-datasets causes the drawback that the samples of the text and image are not initially aligned and this can create slight noise in the training process.

TABLE IV. BASELINE COMPARISON TABLE

Method	Architecture	Dataset	Accuracy	Limitations
Base Paper 1	Large multimodal deep learning architecture	Multi modal dataset	78%	High computational complexity
Base Paper 2	Multimodal emotion detection for e-learning	E-learning dataset	74%	Domain-specific
Proposed Method	DistilBERT + EfficientNet + Gated Fusion	(ISEAR + FER2013 + RAF-DB)	82%	Cross-dataset pairing

Table IV summarizes that, in contrast to multimodal emotion recognition strategies in the past, the given framework proposes the cross-dataset pairing strategy that enables multimodal learning without the need of synchronized multimodal datasets. Secondly, encoders and frozen feature extractors are lightweight, leading to lower computational complexity, and are capable of performing well.

VII. CONCLUSION AND FUTURE WORK

This paper presented a lightweight multimodal emotion recognition framework that integrates textual and visual emotional information to improve emotion classification performance. The proposed system combines contextual text representations extracted using DistilBERT with facial emotion features obtained from EfficientNet-B3. A cross-dataset pairing strategy was introduced to associate textual emotion descriptions from the ISEAR dataset with facial images from the FER2013 and RAF-DB datasets, enabling multimodal learning without requiring synchronised multimodal data. The extracted features are integrated using a gated feature fusion mechanism and classified using an LSTM network. Experimental results demonstrate that the multimodal model achieves superior performance compared with unimodal approaches, highlighting the effectiveness of combining complementary emotional cues from different modalities.

Future work will focus on extending the framework by incorporating additional modalities such as speech or video to further improve emotion recognition accuracy. Advanced fusion strategies based on attention or transformer architectures can also be explored to enhance cross-modal interaction. Furthermore, evaluating the proposed framework on larger multimodal datasets and deploying it in real-time human–computer interaction or healthcare monitoring systems represents promising research directions.

REFERENCES

[1] M. M. Wafa, M. M. Eldefrawi, and M. S. Farhan, "Advancing multimodal emotion recognition in big data through prompt engineering and deep adaptive learning," *Journal of Big Data*, vol. 12, no. 210, 2025, doi: 10.1186/s40537-025-01264-w.

[2] Q. El Maazouzi and A. Retbi, "Multimodal detection of emotional and cognitive states in e-learning through deep fusion of visual and textual data with NLP," *Computers*, vol. 14, no. 314, 2025, doi: 10.3390/computers14080314.

[3] P. 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, 2017, doi: 10.1016/j.inffus.2017.02.003.

[4] Y. Baltrušaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019, doi: 10.1109/TPAMI.2018.2798607.

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, doi: 10.18653/v1/N19-1423.

[6] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," in *Proc. NeurIPS Workshop*, 2019, doi: 10.48550/arXiv.1910.01108.

[7] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, 2019, doi: 10.48550/arXiv.1905.11946.

[8] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, 2022, doi: 10.1109/TAFFC.2020.2981446.

[9] B. Mollahosseini, D. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2019, doi: 10.1109/TAFFC.2017.2740923.

[10] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Information Processing*, 2013, doi: 10.1007/978-3-642-42051-1_16.

[11] A. Zadeh *et al.*, "Tensor fusion network for multimodal sentiment analysis," in *Proc. EMNLP*, 2017, doi: 10.18653/v1/D17-1115.

[12] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S. F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, pp. 3–14, 2017, doi: 10.1016/j.imavis.2017.08.003.

[13] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008, doi: 10.1007/s10579-008-9076-6.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[15] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NeurIPS*, 2017, doi: 10.48550/arXiv.1706.03762.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, doi: 10.48550/arXiv.1412.6980.

[17] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, *A Practical Guide to Sentiment Analysis*. Cham, Switzerland: Springer, 2017, doi: 10.1007/978-3-319-55394-8.