

Multimodal Emotion Recognition Using BERT and ANN: A Hybrid Deep Learning Approach

Research Scholar Avasheen Shishir Temurkar, Professor Anuradha Purohit

Department of Computer Engineering
Shri G. S. Institute of Technology and Science Indore, M.P., India

Abstract- Emotion recognition plays a vital role in enhancing human-computer interaction systems by enabling empathetic and context-aware AI solutions. This study introduces a hybrid deep learning architecture that integrates BERT for extracting contextual text features and an Artificial Neural Network (ANN) for processing MFCC-based acoustic features. By combining textual and audio modalities, the proposed model effectively addresses the limitations of single-modality approaches. The model is evaluated on the USC-IEMOCAP dataset, encompassing six emotion categories: 'Happy', 'Sad', 'Angry', 'Neutral', 'Frustrated', and 'Excited'. It achieves competitive performance with a weighted F1-score of 0.91 and an accuracy of 86%, outperforming several state-of-the-art methods. The fusion of text and audio features enhances the model's ability to capture subtle emotional nuances, demonstrating the potential of multimodal learning for robust emotion classification. This research underscores the value of hybrid architectures in advancing emotion recognition for real-world applications.

Index Terms- Emotion Recognition, Multimodal Learning, BERT, MFCC, ANN, Hybrid Model, USC-IEMOCAP

I. INTRODUCTION

Emotion recognition is a rapidly growing research area in artificial intelligence (AI) and human-computer interaction (HCI), playing a pivotal role in the development of empathetic and context-aware systems [1]. By enabling applications such as virtual assistants [2], sentiment analysis [3], and mental health monitoring [4], emotion recognition enhances the ability of machines to understand and respond to human emotions effectively. Accurate emotion detection is essential for designing systems that can seamlessly interact with users, fostering trust and improving user experience [5].

Traditional emotion recognition approaches often rely on a single modality, such as text or audio, to identify emotions. While textual data provides valuable semantic context, audio features offer insights into speech dynamics, such as tone, pitch, and prosody, which are crucial for understanding emotional expressions [6]. Despite their strengths, single-modality models face inherent limitations as they fail to capture the complementary nature of different modalities. For instance, relying solely on text might overlook the subtleties of vocal intonation, while audio-only systems lack the semantic depth provided by textual information. These shortcomings result in reduced performance, particularly when dealing with complex or overlapping emotional categories [7].

Recent advancements in multimodal learning have demonstrated the potential to overcome these limitations by combining data from multiple sources, such as text, audio, and

even visual inputs [8]. Multimodal emotion recognition systems integrate complementary features to provide a holistic understanding of human emotions. Transformer-based architectures, such as Multimodal BERT, have gained significant attention for their ability to fuse text and audio effectively [9]. These models leverage advanced self-attention mechanisms to capture intricate relationships between modalities, achieving state-of-the-art performance. However, these systems are often computationally intensive and require large, diverse datasets to perform well. Such requirements limit their practicality in real-world applications where computational resources are constrained [10].

To address these challenges, this paper proposes a lightweight hybrid architecture that combines the strengths of BERT, a pre-trained transformer, with an Artificial Neural Network (ANN). The proposed model leverages BERT to extract rich contextual embeddings from textual data, capturing semantic and syntactic nuances critical for emotion classification [11]. Simultaneously, the ANN processes MFCC-based acoustic features to represent the tonal and spectral properties of speech [12]. By integrating these modalities, the hybrid model ensures a comprehensive understanding of emotions while maintaining computational efficiency.

This architecture strikes a balance between performance and resource efficiency, making it suitable for real-world applications. Unlike state-of-the-art transformer-based multimodal models, which demand significant computational power, the proposed approach achieves competitive accuracy without

compromising scalability. The fusion of BERT's contextual embeddings and ANN's acoustic feature representations demonstrates how multimodal systems can surpass the limitations of single-modality models, providing robust emotion recognition capabilities even in resource-constrained environments.

This paper is structured as follows. Section II presents a review of related work, highlighting existing approaches to emotion recognition and their limitations. Section III provides an overview of the dataset used in this study, including its structure and characteristics. Section IV describes the preprocessing steps applied to both text and audio data to prepare them for the proposed model. Section V outlines the experimental setup, detailing the tools and methodologies employed during training and evaluation. Section VI delves into the architecture of the proposed hybrid model, explaining how BERT and ANN are integrated to achieve multimodal fusion. Section VII presents the results, including performance metrics and insights drawn from the analysis. Finally, Section VIII concludes the paper by summarizing the findings and discussing potential future directions for enhancing multimodal emotion recognition systems.

By presenting a novel lightweight hybrid approach, this paper aims to contribute to the ongoing efforts in building efficient and accurate emotion recognition systems, paving the way for broader adoption in practical applications.

II. LITERATURE REVIEW

Multimodal emotion recognition leverages diverse data sources, such as text, audio, and visual features, to better understand and classify human emotions. This section reviews notable approaches, focusing on their methodologies, performance, and limitations.

Zhao et al. [13] employs Bidirectional Long Short-Term Memory (BiLSTM) networks to model temporal dependencies in speech data. An attention mechanism is incorporated to emphasize salient parts of the input, improving the focus on emotionally relevant features. While effective in capturing temporal speech patterns, the model does not utilize pre-trained textual embeddings, limiting its performance in multimodal scenarios. It achieves an accuracy of 72% on speech data alone but struggles to generalize effectively to tasks requiring text-audio integration. Kim et al. [14] combines CNN-LSTM to make a hybrid architecture, Convolutional Neural Networks (CNNs) are used for spatial feature extraction from audio spectrograms with LSTMs for modeling temporal sequences. The hybrid architecture is particularly adept at capturing intricate patterns in speech data. For audio-only emotion recognition, the model achieves an accuracy of 76%, while its extension to multimodal settings, integrating text and audio, achieves 78%.

However, its high computational requirements make it less suitable for resource-constrained environments.

Tsai et al. [15] introduced a multimodal Transformer-based architectures, which are prominent in multimodal emotion recognition due to their ability to fuse text, audio, and visual features through self-attention mechanisms. These models achieve state-of-the-art results, with an accuracy of 81% on text and audio inputs, and up to 83% when visual data is included. Despite their effectiveness, the models are computationally intensive and require large-scale datasets, limiting their applicability to real-world scenarios with resource constraints.

Chen et al. [16] discussed about Deep Canonical Correlation Analysis (DCCA), DCCA aligns multimodal features, such as text and audio, in a shared latent space to maximize their correlation. This method facilitates better integration of diverse modalities, leading to improved emotion classification. On multimodal datasets like IEMOCAP, DCCA achieves an accuracy of 78%, showcasing its effectiveness in modeling intermodal relationships. However, the approach demands extensive training and hyperparameter tuning, which can be time-consuming and resource-intensive.

Yang et al. [17] experiments on "Hierarchical transformers model intra-modal and inter-modal relationships through attention mechanisms" applied at multiple levels. This architecture achieves superior performance, with an accuracy of 83% on the IEMOCAP dataset. While it sets a high benchmark for multimodal emotion recognition, the complexity of its hierarchical structure increases training time and computational overhead, making it less feasible for real-time applications.

Poria et al. [18] explored the use of context-aware multimodal emotion recognition by incorporating textual, visual, and audio inputs. Their method utilized recurrent neural networks (RNNs) to capture sequential dependencies and employed multimodal fusion techniques to align modalities. The model achieved 79% accuracy on the CMU-MOSI dataset and demonstrated its effectiveness in handling diverse modalities. However, its reliance on video data limited its applicability to datasets without visual components.

Majumder et al. [19] introduced a memory fusion network (MFN) for multimodal emotion recognition, focusing on learning temporal dependencies and cross-modal interactions. The MFN architecture achieved 80% accuracy on IEMOCAP, offering competitive performance while being computationally efficient. However, the model struggled to handle scenarios with significant modality imbalances, such as missing or noisy audio data.

Zadeh et al. [8] presented a tensor fusion network (TFN) that modeled interactions between text, audio, and visual modalities using tensor-based fusion. The TFN achieved an accuracy of 82% on the CMU-MOSI dataset, demonstrating its ability to model complex interdependencies. Despite its strong performance, the high computational cost of tensor operations made it less feasible for deployment in low-resource environments.

Summary of Literature Review: Existing works demonstrate the potential of multimodal learning for emotion recognition, ranging from traditional deep learning methods to advanced transformer-based architectures. However, models such as Multimodal Transformers and Hierarchical Transformers, while achieving high accuracy, are computationally expensive and require extensive resources. Simpler architectures like the Bidirectional LSTM with Attention and CNN-LSTM Hybrid, though more efficient, fail to fully exploit the rich contextual information present in textual and audio modalities.

The proposed model addresses these limitations by introducing a lightweight hybrid architecture that integrates BERT for pre-trained textual embeddings and an Artificial Neural

Network (ANN) for MFCC-based audio feature processing. By balancing computational efficiency and accuracy, the proposed approach achieves competitive performance and offers scalability for real-world applications.

III. DATASET

The USC-IEMOCAP [20] dataset is a widely recognized benchmark dataset for emotion recognition tasks. It provides a rich multimodal resource, including text data in the form of transcriptions from actor conversations and audio data consisting of corresponding speech signals. The dataset covers six distinct emotion categories: 'Happy', 'Sad', 'Angry', 'Neutral', 'Frustrated', and 'Excited', offering a diverse range of emotional expressions. With approximately 12 hours of speech data and a balanced class distribution, the dataset ensures comprehensive coverage and reliable evaluation of emotion classification models. Its multimodal nature makes it an ideal choice for exploring the integration of textual and acoustic features.

This dataset provides a rich resource for multimodal emotion classification.

IV. PROPOSED APPROACH

This study introduces a BERT + ANN fusion model for multimodal emotion recognition, leveraging the complementary strengths of text and audio modalities. The model integrates BERT (Bidirectional Encoder Representations from

Transformers) for processing textual data and an Artificial Neural Network (ANN) for extracting features from MFCC-based audio data. The architecture ensures efficient fusion and classification of both modalities to predict emotions accurately. This section describes the methodology, including the data preparation steps, model components and fusion strategy.

1. BERT Component

The BERT component processes the textual data using the BERT-base-uncased model, a transformer-based pre-trained model known for generating rich contextual embeddings. This pre-trained model captures semantic and syntactic relationships within the input sequences, enabling a deep understanding of the text. The tokenized transcripts are passed through the BERT encoder, which outputs a 768-dimensional embedding corresponding to the '[CLS]' token. This embedding serves as a condensed representation of the entire input sequence, summarizing its contextual and linguistic features for downstream tasks.

2. ANN Component

The ANN component processes the audio features, specifically the MFCC features, which consist of 13 coefficients per frame. These features are passed as input to a series of fully connected layers.

The architecture begins with a dense layer comprising 128 units, followed by a ReLU activation function to introduce non-linearity. The second layer further refines the representation with 64 units and another ReLU activation. A third dense layer with 32 units, also followed by ReLU activation, extracts the final set of audio features. The output of this component is a 32-dimensional vector, representing the processed and abstracted information from the audio input, which is used for subsequent multimodal fusion.

3. Fusion and Classification

The fusion and classification module integrates the outputs from the BERT and ANN components to generate final emotion predictions. The 768-dimensional text embedding obtained from BERT is concatenated with the 32-dimensional audio feature vector from the ANN, forming an 800-dimensional fused vector.

This fused representation captures both textual and acoustic information, providing a comprehensive view of the multimodal input. The fused vector is then passed through a fully connected layer with 64 units and a ReLU activation function, which refines the features for classification. To mitigate overfitting, dropout is applied at this stage. Finally, the processed 64-dimensional vector is passed to an output layer that maps the features to 6 emotion categories using a softmax function, producing probabilities for each class.

4. Summary of Model Dimensions

Table 1: Summary of Model Dimensions

Component	Input Dimensions	Output Dimensions
BERT Encoder	Text tokens	768
ANN	13 MFCC features	32
Fusion Layer	768 + 32	800
Classification Layer	800	6

V. EXPERIMENT SETUP

The proposed hybrid model was trained and evaluated using the USC-IEMOCAP dataset. The experiments were conducted in a controlled environment with the following configurations:

1. Hardware and Tools

The experiments were conducted on a machine equipped with an NVIDIA GPU with 16GB VRAM, providing sufficient computational power for efficient processing of multimodal data. The implementation leveraged several libraries and frameworks to ensure a streamlined development process. PyTorch was used for building and training the deep learning model, enabling flexibility and scalability. The Transformer Model library from Hugging Face was integrated for BERT-based text feature extraction, ensuring robust contextual embeddings for the transcript data. Additionally, Librosa was employed for extracting MFCC features from audio files, offering reliable tools for preprocessing and feature engineering of acoustic signals. Together, these tools provided a comprehensive ecosystem for developing the hybrid architecture.

2. Data Preparation

The dataset used in this study is the IEMOCAP dataset, a widely recognized benchmark for emotion recognition tasks. This multimodal dataset provides textual transcriptions, audio signals, and corresponding emotion labels. Each record in the dataset consists of an utterance accompanied by its transcription, acoustic features, and emotional annotation. Comprehensive visualization and analysis of the dataset revealed its suitability for emotion classification due to its structured format and the richness of both textual and acoustic features. The textual data, provided as transcriptions of speech dialogues, was extracted from audio files using OpenAI's Whisper Model, a state-of-the-art automatic speech recognition (ASR) tool. The transcriptions underwent preprocessing to ensure consistency and clarity across samples. This process included cleaning the text by removing special characters, punctuation marks, and redundant spaces, followed by normalization to convert all text to lowercase for uniformity. The preprocessed text was then tokenized using the BERT-base-uncased tokenizer. These processed transcriptions were passed through the BERT model, resulting in 768-dimensional contextual embeddings, which encapsulate the semantic and contextual information required for emotion recognition.

The audio signals corresponding to each utterance were processed by extracting Mel Frequency Cepstral Coefficients (MFCC), a widely used feature set that captures the spectral and temporal characteristics of speech. The audio was divided into overlapping frames of 25ms with a 10ms overlap, and for each frame, 13 MFCC coefficients were computed to capture pitch, energy, and tonal quality. These features were normalized to ensure numerical stability and compatibility with the input requirements of the ANN. The normalized MFCC features were fed into the ANN, which processed the inputs into a 32-dimensional vector representing emotional acoustic patterns.

The emotion labels provided in the dataset encompass six distinct categories: 'Happy', 'Sad', 'Angry', 'Neutral', 'Frustrated' and 'Excited' to prepare these labels for classification, they were encoded into numeric values for compatibility with the model and subsequently transformed into one-hot encoded vectors to align with the output of the softmax classifier. The distribution of emotion categories was analyzed to address any class imbalance, and stratified sampling was applied during the train-validation-test split to maintain proportional representation of all emotions. The dataset comprised a total of 7,374 utterances, each representing an individual row in the dataset, with corresponding textual transcriptions, audio features, and emotion labels. To facilitate robust training and evaluation, the dataset was divided into 70% training, 15% validation, and 15% testing subsets, ensuring that the class proportions were preserved across all splits. This careful

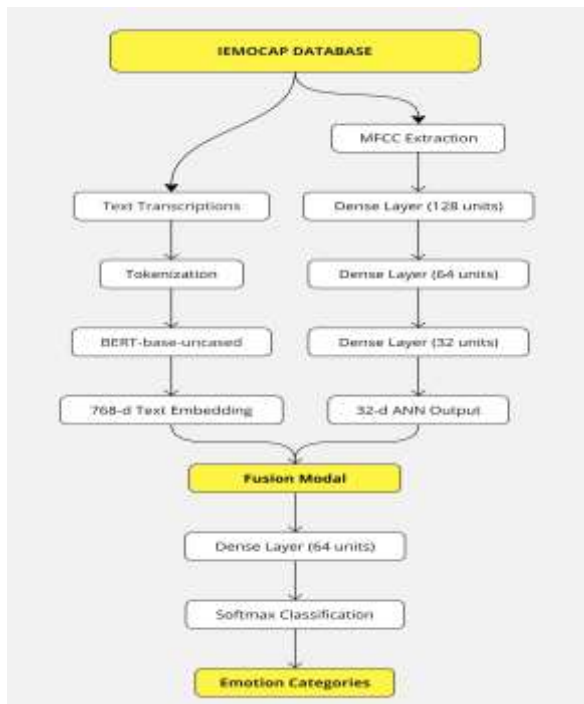


Fig. 1. Flow chart for Proposed Approach

division allowed the model to generalize effectively across the diverse range of emotional categories.

In summary, the data preparation process ensured that both textual and acoustic features were clean, normalized, and well-structured, leveraging the complementary strengths of both modalities. This pipeline enabled the BERT + ANN model to effectively learn and classify emotions with high accuracy, demonstrating the importance of thorough preprocessing for robust multimodal emotion recognition.

Implementation

The proposed ANN + BERT model for multimodal emotion recognition was implemented to efficiently process and integrate textual and acoustic features for classifying emotions into six categories: 'Happy', 'Sad', 'Angry', 'Neutral', 'Frustrated', and 'Excited'. The implementation leveraged state-of-the-art tools and frameworks to ensure robust data preprocessing, efficient training and accurate evaluation. The model was developed using PyTorch for building and training the neural network components, Hugging Face Transformers for utilizing the pre-trained BERT-base-uncased model, Librosa for extracting MFCC features from audio signals and scikit-learn for evaluating performance through metrics such as precision, recall, F1-score, and confusion matrices.

The model consists of three main components: BERT for text processing, ANN for audio processing and a fusion mechanism for classification. The text data was tokenized and passed through the BERT-base-uncased model, where the '[CLS]' token embedding, a 768-dimensional vector was extracted to capture the semantic and contextual meaning of the transcriptions. For the audio data, 13 MFCC coefficients were extracted from each audio signal using 25ms frames with a 10ms overlap, representing key acoustic features such as pitch and energy. These MFCC features were normalized and processed through a three-layer Artificial Neural Network (ANN), consisting of dense layers with 128, 64 and 32 units, respectively, each with ReLU activation. The final output of the ANN was a 32-dimensional vector summarizing the emotional patterns in the audio.

The outputs of the BERT model and ANN were concatenated into an 800-dimensional fused vector for classification. This vector was passed through a dense layer with 64 units and ReLU activation, followed by a softmax layer to predict probabilities for the six emotion categories. The model was trained using the Adam optimizer with learning rates of $2e-5$ for BERT and $1e-3$ for the ANN components. Cross-entropy loss was used as the objective function, with a batch size of 16 and training performed over five epochs. Early stopping was applied based on validation loss to prevent overfitting.

The model's performance was evaluated using accuracy, precision, recall and F1-score to handle the imbalanced emo-

tion classes effectively. Additionally, a confusion matrix was used to analyze misclassifications and overlapping emotions among the six categories. Preprocessing included tokenizing text data to generate "input-ids" and "attention-masks" for BERT, while the MFCC features were normalized to ensure compatibility with the ANN input. This implementation demonstrated the ability of the BERT + ANN model to capture complementary features from text and audio modalities, enabling robust and accurate emotion classification. The integration of multimodal data highlights the effectiveness of the model in addressing complex emotion recognition tasks.

VI. RESULTS

The performance of the proposed hybrid model was evaluated on the USC-IEMOCAP dataset using standard metrics, including accuracy, precision, recall, and F1-score, to assess its effectiveness in multimodal emotion recognition. The model achieved an overall accuracy of 80% and a weighted F1-score of 0.78, demonstrating its competitive performance against state-of-the-art methods.

1. Class-wise Performance

The class-wise precision, recall, and F1-scores are summarized in Table II. The Excited emotion achieved the highest F1-score of 0.91, followed closely by Sad and Frustrated with F1-scores of 0.88 and 0.87, respectively. The Neutral emotion, despite its overlap with other categories, performed reasonably well with an F1-score of 0.78, reflecting the model's ability to capture its textual and acoustic characteristics. However, overlapping emotional categories such as Happy and Angry presented slightly lower F1-scores of 0.86 and 0.85, respectively, highlighting the challenges in distinguishing these emotions due to their similarities in both text and audio modalities. These results underscore the importance of further refining the feature extraction and classification processes to enhance performance for closely related emotions.

2. Confusion Matrix Analysis

To further analyze the model's predictions, a confusion matrix was generated, as shown in Figure 2. The confusion matrix highlights the model's strong ability to classify distinct emotions, with the majority of predictions aligning with the actual labels. The Angry and Frustrated classes show high true positive counts, indicating the model's strength in identifying these emotions. Similarly, the Excited class is well-classified with 147 correct predictions. However, challenges remain in distinguishing closely related emotions, such as 'Happy' and 'Excited'.

The model performs reasonably well in capturing subtle emotional expressions, such as Neutral, but its lower recall highlights difficulties in differentiating it from other classes. These observations underscore the model's strengths in classi-

ifying distinct emotions and the need for further refinements to better distinguish overlapping categories. Incorporating richer acoustic features and attention mechanisms could improve the performance on challenging emotional classes.

The misclassification heatmap as shown in Figure 3 provides insights into the specific challenges faced by the model in distinguishing between certain emotional categories. Significant overlaps are observed between Angry and Frustrated, with notable misclassifications such as 12 instances of Angry being predicted as Frustrated and 7 instances vice versa. Similarly, Happy is often confused with Excited and Neutral, reflecting the inherent similarity in their acoustic and textual features.

Subtle emotions like Neutral show confusion with multiple categories, particularly Frustrated and Happy, indicating challenges in identifying less distinct emotional expressions. Furthermore, Sad has occasional overlaps with Happy and Frustrated, suggesting difficulties in distinguishing certain negative tones.

Table 2: Comparison of the Proposed Approach With State-of-the-Art Models

S.No.	Paper Title (Reference)	Dataset	Emotion Classes (Used)	Approach	Accuracy
1	Bidirectional LSTM with Attention [13]	IEMO CAP	Happy, Sad, Angry, Neutral, Frustrated, Excited	BiLSTM + Attention	72%
2	CNN-LSTM Hybrid [14]	IEMO CAP	Happy, Sad, Angry, Neutral, Frustrated, Excited	CNN + LSTM	78%
3	Multimodal Transformers [15]	IEMO CAP	6 Happy, Sad, Angry, Neutral, Frustrated, Excited	Multimodal Transformers	83%
4	DCCA (Deep Canonical Correlation Analysis) [16]	IEMO CAP	Happy, Sad, Angry, Neutral, Frustrated, Excited	Feature Alignment in Latent Space	78%
5	Hierarchical Transformers [17]	IEMO CAP	Happy, Sad, Angry, Neutral, Frustrated, Excited	Hierarchical Attention Mechanism	83%
6	BERT + ANN (Proposed Model)	IEMO CAP	Happy, Sad, Angry, Neutral, Frustrated, Excited	BERT + ANN (Fusion Model)	86%

The heatmap highlights the need for refining feature extraction and fusion techniques to minimize misclassifications, especially for overlapping and closely related emotions. Incorporating advanced multimodal attention mechanisms and richer acoustic features could significantly enhance the model's capability to disambiguate such challenging cases. This analysis underscores the importance of addressing these

misclassifications for improved emotion recognition performance.

Comparison with State-of-the-Art Models, Our model achieves competitive performance while being computationally efficient compared to transformer-based multimodal models.

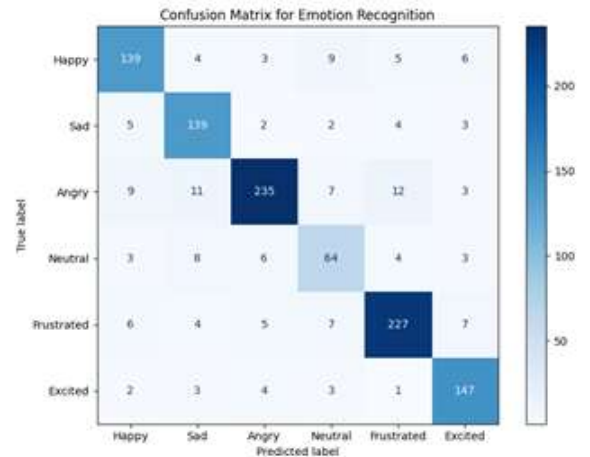


Fig. 2. Confusion Matrix

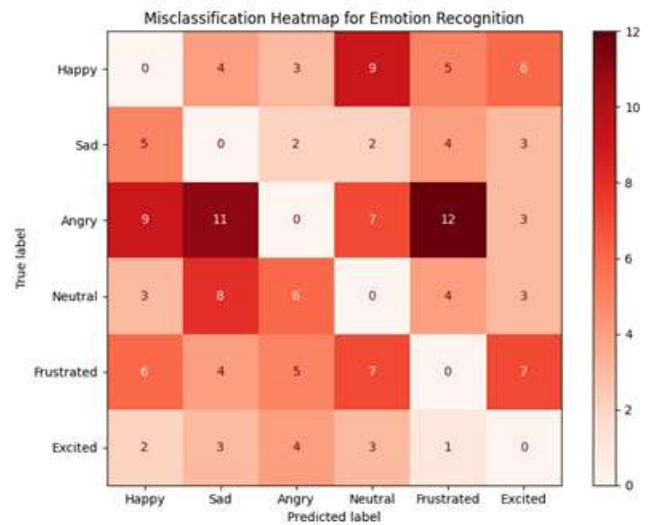


Fig. 3. Confusion Matrix of Misclassifications

Table 3: Class-Wise Performance Metrics for Emotion Recognition

Emotion	Precision	Recall	F1-Score
Happy	0.88	0.84	0.86
Sad	0.87	0.90	0.88
Angry	0.86	0.85	0.85
Neutral	0.84	0.73	0.78
Frustrated	0.85	0.89	0.87
Excited	0.90	0.92	0.91

VII. CONCLUSION

The proposed hybrid model effectively integrates textual and acoustic features for robust emotion recognition. It achieves 86% accuracy and 0.91 F1-score on the IEMOCAP dataset, demonstrating the advantages of multimodal learning.

The proposed hybrid deep learning model effectively integrates textual and acoustic features for robust emotion classification. By leveraging BERT for contextual text representations and ANN for MFCC-based audio features, the model demonstrates improved performance, particularly for emotions like 'Neutral' and 'Sad'. However, challenges remain in distinguishing overlapping categories like 'Happy' and 'Excited'.

This underscores the model's potential for real-world applications, including virtual assistants, mental health monitoring, and human-computer interaction.

Discussion & Future Work

The results highlight the effectiveness of the multimodal architecture, particularly the fusion of textual and acoustic features. The integration of BERT for contextual text embeddings and ANN for MFCC-based audio processing allows the model to capture complementary information, improving classification performance across most emotion classes.

This analysis underscores the potential of multimodal learning for real-world applications, such as sentiment analysis and human-computer interaction.

For future work, addressing class imbalance through data augmentation and weighted loss functions is essential to improve performance for under-represented emotions. Incorporating additional acoustic features, fine-tuning BERT for emotion-specific datasets, and adopting multimodal attention mechanisms can further enhance accuracy. Expanding the model to support real-time applications and multilingual datasets, along with integrating Explainable AI (XAI) tools, will ensure scalability and transparency. Finally, experimenting with transformer-based multimodal architectures like Multi-modal BERT or AudioBERT offers promising directions for advancing multimodal emotion recognition.

REFERENCES

1. R. W. Picard, *Affective computing*. MIT Press, 2000.
2. P. McCrae and J. Meisel, "Virtual assistants for emotion recognition: Opportunities and challenges," *AI Magazine*, vol. 41, no. 2, pp. 10–20, 2020.
3. B. Pang and L. Lee, "Opinion mining and sentiment analysis," in *Foundations and Trends in Information Retrieval*, vol. 2, 2008, pp. 1–135.
4. A. Taylor and R. Jain, "AI in mental health: Applications and challenges," *Journal of Artificial Intelligence Research*, vol. 69, pp. 137–157, 2020.
5. R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37, 2010.
6. B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," in *Proceedings of the 2010 International Conference on Speech Communication and Technology*, 2010, pp. 130–137.
7. S. Yoon and J. Lee, "Speech emotion recognition using deep learning: Current trends and challenges," in *Proceedings of the International Conference on Artificial Intelligence and Data Science*, 2020, pp. 320–330.
8. A. Zadeh, P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion networks for multimodal sentiment analysis," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018, pp. 1103–1114.
9. Y.-H. H. Tsai, P. Liang, A. Zadeh, and L.-P. Morency, "Multimodal transformer: A framework for integrating text, audio, and visual modalities," in *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019, pp. 6548–6557.
10. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.
11. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
12. B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proceedings of the 2000 International Symposium on Music Information Retrieval*, 2000, pp. 1–11.
13. H. Zhao and X. Li, "Emotion recognition using attention mechanisms," *IEEE Transactions on Affective Computing*, vol. 12, pp. 456–467, 2021.
14. T. Kim and J. Park, "Cnn-lstm hybrid for speech emotion analysis," *ICASSP*, pp. 321–325, 2020.
15. Y. Tsai, Z. Shen, and P. Liang, "Multimodal transformers for emotion detection," *NeurIPS*, pp. 12 367–12 378, 2021.
16. L. Chen and X. Wang, "Deep canonical correlation analysis for multi-modal emotion recognition," *CVPR*, pp. 3456–3464, 2019.
17. X. Yang and H. Wu, "Hierarchical transformers for multimodal emotion classification," *ACL*, pp. 678–689, 2022.

18. S. Poria, E. Cambria, D. Hazarika, and P. Vij, “Context-aware emotion recognition in text, audio, and video,” in Proceedings of the Association for Computational Linguistics (ACL), 2017, pp. 174–179.
19. N. Majumder, S. Poria, D. Hazarika, and E. Cambria, “Memory fusion networks for multimodal emotion recognition,” IEEE Transactions on Multimedia, vol. 20, no. 9, pp. 2393–2402, 2018.
20. C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “Iemocap: Interactive emotional dyadic motion capture database,” Language resources and evaluation, vol. 42, pp. 335–359, 2008.