International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 4, July-Aug-2026, ISSN (Online): 2395-566X

A BiGRU-Based Multimodal Framework for Emotion Recognition Using Text, Speech, and Conversational Context

Abirami A[1], Miranda Lakshmi T[2], Martin A[3]

[1](Assistant Professor & Research Scholar PG & Research Department of Computer Science and Artificial Intelligence, St. Joseph's College of Arts and Science (Autonomous), Cuddalore, Affiliated to Annamalai University, Tamilnadu abi.mail87@gmail.com)

2(Associate Professor PG & Research Department of Computer Science and Artificial Intelligence, St. Joseph's College of Arts and Science (Autonomous),Cuddalore, Affiliated to Annamalai University, Tamilnadu drtmirandalakshmi@gmail.com)

3(Assistant Professor Department of Computer Science, Central University of Tamil Nadu,Thiruvarur 610005, India martin@cutn.ac.in)

Abstract — Conversational emotion recognition is inherently challenging because an utterance's emotional label does not depend on wording alone — it is shaped jointly by content, delivery, and the dialogue history surrounding it. This work introduces a multimodal deep-learning pipeline that fuses textual, acoustic, and contextual signals to classify emotion at the utterance level within conversations. Linguistic meaning is captured through pretrained BERT embeddings, broad acoustic patterns through Wav2Vec2 embeddings, and explicit vocal-affect cues through handcrafted prosodic descriptors (pitch statistics and short-time energy). Once concatenated at the utterance level, these heterogeneous features are passed into a Bidirectional Gated Recurrent Unit (BiGRU) that captures dependencies across dialogue turns, with a fully connected head performing the final classification. Because the benchmark dataset's seven-class emotion taxonomy is heavily skewed, a weighted cross-entropy loss built from inverse class-frequency statistics is applied to offset this imbalance during training. On the MELD (Multimodal EmotionLines Dataset) benchmark, the proposed pipeline reaches 77.56% overall accuracy and a weighted F1-score of 0.804, with per-class ROC-AUC spanning 0.89 to 0.96. Confusion-matrix and error analysis reveal that common emotions such as neutral and joy are classified reliably, whereas rarer emotions such as fear and disgust remain substantially harder to recognize even with class-weighted training — underlining how persistent the data-imbalance problem remains in this domain and pointing toward future work on stronger imbalance-handling and context-modelling techniques.

Keywords— Multimodal Emotion Recognition, Emotion Recognition in Conversation (ERC), BERT, Wav2Vec2, Bidirectional GRU, MELD Dataset, Class Imbalance, Conversational AI, Deep Learning .

I. INTRODUCTION

Interpreting emotion from natural conversation sits at the core of affective computing, with uses ranging from empathetic virtual assistants and mental-health monitoring to call-centre analytics, opinion mining, and social robotics. Unlike single-sentence sentiment analysis, Emotion Recognition in Conversations (ERC) has to reckon with the fact that an utterance's emotional label is shaped not just by its own wording but also by the dialogue context around it, the speaker's vocal tone, how turns are exchanged, and the emotional arc of the conversation overall. Because spoken dialogue conveys meaning through both words and vocal delivery at once, this context-dependence makes ERC considerably harder than conventional text-only sentiment classification.

Powerful pretrained encoders for language and speech — BERT for text and Wav2Vec2 for raw audio have emerged from recent progress in self-supervised representation learning, and both can be repurposed as feature extractors for emotion classification without any task-specific pretraining [2][3]. Combining these two very different modalities effectively, while also carrying contextual information across a multi-turn dialogue, is still an unresolved design problem. Systems relying on text alone lose paralinguistic signals pitch, loudness that frequently reveal sarcasm, sincerity, or emotional intensity, while acoustic-only systems lose the lexical grounding that anchors an utterance's meaning. A well-rounded ERC system consequently needs both modalities together with an explicit

way of tracking how emotional state shifts over the course of a conversation.

A separate practical difficulty is that widely used conversational emotion datasets — MELD chief among them suffer from pronounced class imbalance [1]: because everyday conversation is mostly neutral in tone, neutral utterances vastly outnumber emotionally charged ones like fear or disgust. When trained without accounting for this, models tend to default toward the majority class, producing accuracy figures that look strong on paper while failing on precisely the minority emotions that matter most in practice.

This paper proposes a multimodal framework built around three utterance-level components: BERT-derived textual embeddings, Wav2Vec2-derived acoustic embeddings, and handcrafted prosodic features (pitch statistics and energy). For every dialogue, the fused multimodal representation of each utterance flows through a Bidirectional GRU [4] that captures contextual dependencies across the conversation, after which a fully connected classification head assigns an emotion label to each utterance. Class imbalance is addressed during training via a weighted cross-entropy loss using inverse-frequency class weights. The framework is evaluated on the MELD benchmark using accuracy, per-class precision/recall/F1, confusion-matrix analysis, and ROC-AUC, alongside a detailed look at error patterns and limitations. The contributions of this work can be summarized as follows:

- A unified multimodal pipeline combining pretrained text (BERT) and speech (Wav2Vec2) encoders with lightweight prosodic descriptors (pitch, energy) for rich utterance-level emotion representation.
- A BiGRU-based contextual sequence model that captures emotional dependencies and conversational flow across dialogue turns, operating directly on the fused multimodal utterance embeddings.
- A class-imbalance-aware training strategy using inverse-frequency weighted cross-entropy [6][12][17], evaluated comprehensively on the MELD dataset using accuracy, macro/weighted F1, confusion matrix, per-class ROC-AUC, and prediction-confidence analysis.
- A detailed empirical error analysis identifying which emotion pairs are most frequently confused and discussing the underlying causes, together with concrete directions for future improvement.

The rest of this paper unfolds as follows. Section II surveys related work spanning text-based, speech-based, and multimodal ERC, class-imbalance techniques, and a comparative look at existing MELD methods. Section III lays out the proposed methodology — feature extraction, fusion, and the contextual sequence model. Section IV covers the experimental setup along with results and discussion. Section V examines this study's limitations and threats to validity. Section VI turns to computational complexity, reproducibility, and ethical considerations. Section VII closes the paper and points to future work.

II. LITERATURE SURVEY

Text-Based Emotion and Sentiment Recognition

Text-based emotion recognition originally relied on lexicon-driven and classical machine-learning approaches built on hand-engineered n-gram and part-of-speech features. That landscape shifted considerably once pretrained transformer language models arrived. BERT, put forward by Devlin et al., is a bidirectional transformer pretrained through masked language modelling and next-sentence prediction; it yields contextualized token- and sentence-level embeddings that transfer well across many downstream NLP tasks — emotion and sentiment classification included — whether through light fine-tuning or as a frozen extractor [2]. The self-attention mechanism at BERT's core traces back to Vaswani et al.'s Transformer architecture, which replaced recurrence altogether with parallelizable self-attention for modelling sequences [7]. Within ERC specifically, text-only baselines built on BERT or comparable encoders hold up reasonably well but are consistently beaten by models that also draw on acoustic and contextual signal, as the original MELD benchmark study demonstrates [1].

Speech Emotion Recognition

Handcrafted acoustic descriptors — Mel-Frequency Cepstral Coefficients, pitch-contour statistics, and energy — formed the backbone of earlier speech emotion recognition (SER) systems, typically paired with classifiers such as SVMs or HMMs. Extracting pitch (fundamental frequency, F0) is usually done through autocorrelation-style algorithms; here we use the YIN algorithm from de Cheveigné and Kawahara, which relies on a cumulative mean normalized difference function to yield robust F0 estimates for speech and music [8]. Handcrafted features have since given way, in many state-of-the-art SER systems, to self-supervised acoustic encoders. Wav2Vec2 is one such encoder — a convolutional-transformer trained through contrastive self-supervised learning directly on raw waveforms, yielding representations that capture phonetic and paralinguistic information relevant to emotion without needing labelled audio during pretraining [3]. We use Wav2Vec2 as our acoustic encoder but keep explicit pitch and energy descriptors alongside it, treating the two feature families as complementary rather than redundant.

Multimodal and Context-Aware Emotion Recognition in Conversations The Multimodal EmotionLines Dataset (MELD) extends the text-only EmotionLines corpus with synchronized audio and visual modalities drawn from the TV series Friends, and introduces several multimodal baselines that establish the importance of both dialogue context and modality fusion for ERC [1]. IEMOCAP, contributed by Busso et al., is another frequently used multimodal conversational benchmark, consisting of roughly 12 hours of scripted and improvised two-party actor conversations labelled with categorical and dimensional emotion annotations [16]. A large body of follow-up work has built increasingly elaborate context encoders on top of these benchmarks. DialogueRNN, from Majumder et al., tracks conversational emotion dynamics through several interacting recurrent networks that separately model global context, speaker (party) state, and emotional state, guided by an attention mechanism over prior utterances [9]. DialogueGCN, from Ghosal et al., takes a different approach, representing a dialogue as a graph of utterances and applying graph convolutions to capture both nearby and distant contextual relationships [10]. Across this line of work, context-aware and multimodal architectures consistently beat context-agnostic, unimodal baselines on MELD, though minority emotions such as fear and disgust remain difficult to recognize because of severe data imbalance [1][5]. Compared with these graph- and attention-heavy context encoders, this work opts for a lighter-weight Bidirectional GRU; the GRU itself was introduced by as a computationally efficient gated-recurrent alternative to the earlier Long Short-Term Memory (LSTM) network of Hochreiter and Schmidhuber [4][11].

Class imbalance shows up throughout emotion recognition datasets, where neutral or mildly positive utterances tend to dominate. Cost-sensitive learning — reweighting the loss so minority-class errors carry a heavier penalty is a standard countermeasure, and it is the strategy used here, via inverse-frequency ('balanced') class weights inside a standard cross-entropy loss [6]. Lin et al. proposed a different remedy from the object-detection literature, Focal Loss, which suppresses the contribution of easy, already well-classified examples so that training concentrates on harder ones [12]. A third, data-level strategy is synthetic oversampling — Synthetic Minority Over-sampling Technique (SMOTE) creates new minority-class samples by interpolating between existing ones in feature space [17].

Comparative Overview of Existing ERC Methods on MELD
Table 1 places the design decisions behind this work in the context of the broader ERC literature, listing representative methods evaluated on MELD in roughly increasing order of context-modelling sophistication, alongside the weighted F1-scores each publication reports. Earlier context-aware models such as bc-LSTM [21] and DialogueRNN [9] lean on recurrent architectures to pass information between utterances; graph-based approaches such as DialogueGCN [10] instead treat utterance dependencies explicitly as a graph; and the strongest results come from more recent transformer-based methods — fine-tuned BERT/RoBERTa baselines and the commonsense-augmented COSMIC framework [22] — which pair large pretrained language models with dialogue-level context.

Handling Class Imbalance

Method	Context Modelling Approach	Reported Weighted F1 on MELD (%)	Ref.
bc-LSTM	Bidirectional LSTM over utterances (unimodal text)	55.9	[21]
DialogueRNN	Party-state and global-context tracking RNN with attention	57.0	[9]
DialogueGCN	Graph convolutional network over utterance graph	58.1	[10]
Fine-tuned BERT	Transformer encoder, context-agnostic	60.3	[2]
RoBERTa-based (fine-tuned)	Transformer encoder with dialogue-level fine-tuning	63.2	[20]
COSMIC	Commonsense-augmented RoBERTa features + GRU	65.2	[22]
Proposed (this work)	BERT + Wav2Vec2 + prosodic features + BiGRU	80.4*	—

Table 1: Comparison with Representative ERC Methods on MELD (Weighted F1)

A caveat worth stating plainly: the 80.4% weighted F1-score reported for this work is not directly comparable to the literature figures above, which come from the official MELD train/dev/test split evaluated with masked per-utterance scoring. This study instead uses a custom 80/20 dialogue-level split with zero-padded shorter dialogues, and, as Section V explains, padded positions are not masked out when computing metrics — which likely inflates the effective weighted F1 relative to official-split figures by counting a larger apparent share of trivially-classified padded (neutral-labelled) positions. Table 1's purpose, then, is to situate this work's design choices and give a rough sense of relative difficulty across methods on MELD, not to assert a state-of-the-art result under a directly comparable protocol.

III. METHODOLOGY

Figure 1 illustrates the overall architecture. For a dialogue $D = \{u_1, u_2, \ldots, u_n\}$ made up of n ordered utterances, each utterance u_i passes independently through a text branch and an audio branch. The two resulting modality-specific embeddings are concatenated per utterance to produce a fused representation x_i, and the resulting sequence $\{x_1, \ldots, x_n\}$ for the dialogue is processed by a Bidirectional GRU that carries conversational context forward, with a fully connected classification head ultimately producing an emotion label $\hat{y}_i$ for each utterance.

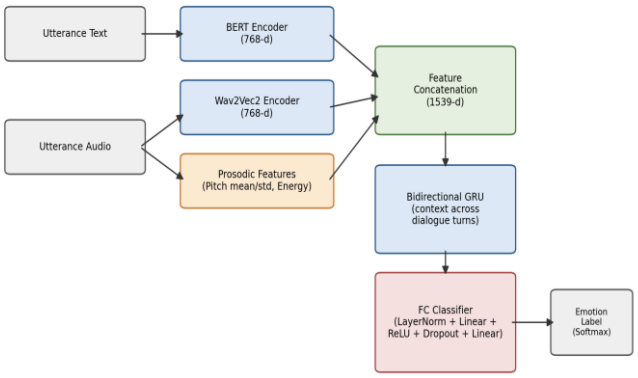

Figure 1: Proposed multimodal emotion recognition Architecture

Textual Feature Extraction

Utterance transcripts are tokenized with the BERT WordPiece tokenizer and fed through a pretrained BERT-base encoder (12 transformer layers, 768 hidden units). Mean-pooling the final layer's token-level hidden states across the sequence dimension yields a fixed 768-dimensional sentence embedding t_i that reflects the utterance's semantic and lexical content regardless of length.

Acoustic Feature Extraction

Each utterance's corresponding audio segment is loaded, downmixed to mono where needed, and resampled to 16 kHz before being passed through a pretrained Wav2Vec2-base model; mean-pooling the final layer's frame-level hidden states over time gives a 768-dimensional acoustic embedding a_i. Alongside this deep embedding, low-level prosodic descriptors are extracted directly from the waveform: pitch (fundamental frequency) is tracked via the YIN algorithm [8] across a 50–300 Hz range, from which mean and standard deviation are computed, while short-time energy is taken as the mean root-mean-square (RMS) amplitude across frames using the librosa library [13]. Together these form a compact 3-dimensional prosodic vector $p_i = [\mu^R_0, \sigma^R_0, \text{RMS}]$, adding explicit pitch-height, pitch-variability, and loudness signals known in the speech-emotion literature to track arousal and valence, on top of the deep acoustic embedding. When audio decoding fails for an utterance, both the acoustic embedding and prosodic vector default to zero, letting the pipeline continue rather than dropping the utterance.

Multimodal Fusion and Contextual Modelling

The 768-dimensional text embedding, 768-dimensional audio embedding, and 3-dimensional prosodic vector are concatenated into one 1539-dimensional utterance representation $x_i = [t_i ; a_i ; p_i]$. Per dialogue, this sequence of fused utterance-level vectors is passed into a single-layer Bidirectional GRU with a hidden size of 128 in each direction. The forward and backward GRU cells update their hidden state at every time step following the standard gated-recurrent-unit formulation from Cho et al. [4]:

$$z_t = \sigma(W_\gamma x_t + U_\gamma h_{t-1}), \quad r_t = \sigma(W_r x_t + U_r h_{t-1})$$

$$\tilde{h}_t = \tanh(W_h x_t + U_h (r_t \odot h_{t-1})), \quad h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$

Here z_t and r_t denote the update and reset gates, and $\odot$ represents element-wise multiplication. At every time step, the forward and backward hidden states are concatenated into a 256-dimensional context-aware representation per utterance. These contextualized states then pass through a classification head made up of: (i) layer normalization over the 256-dimensional GRU output [19]; (ii) a linear layer projecting down to 128 dimensions; (iii) a ReLU activation; (iv) dropout at probability 0.3 for regularization [18]; and (v) a final linear layer mapping to the seven emotion classes, with a softmax producing class probabilities at inference time.

Handling Class Imbalance

MELD skews heavily toward the neutral class, as Table 3 quantifies. To counter the resulting bias toward majority-class predictions, per-class weights w^c are computed with the inverse-frequency ('balanced') formula $w^c = N / (K \cdot n^c)$ — N being the total labelled-utterance count, K the number of classes, and n^c the utterance count for class c. A further scaling

factor of 1.2 is applied to these weights before folding them into a weighted cross-entropy loss, which pushes the model to penalize misclassifications of minority classes like fear and disgust more heavily during training.

Training Configuration

Implemented in PyTorch [15], the model is trained end-to-end while keeping the pretrained BERT and Wav2Vec2 encoders frozen in a feature-extraction-only role only the BiGRU and classification head receive updates during training. Optimization relies on the Adam optimizer from Kingma and Ba [14], using a 2×10^{-4} learning rate, batches of 8 dialogues, and 30 epochs. Dialogue sequences within a batch are padded out to that batch's longest dialogue, and predictions are flattened across time steps when computing the loss. In total, the trainable pipeline holds approximately 1.32 million parameters.

IV. ANALYSIS

Experimental Setup

The MELD benchmark [1] — roughly 13,000 multi-party conversational utterances drawn from 1,433 dialogues in the TV series Friends, labelled across seven emotion categories (neutral, joy, anger, sadness, fear, disgust, surprise) — forms the basis for all experiments. Utterances are kept grouped by dialogue to preserve conversational structure, with an 80/20 train–test split applied at the dialogue level. Evaluation covers overall accuracy, per-class precision/recall/F1-score, macro- and weighted-averaged F1, confusion-matrix analysis, per-class ROC-AUC (one-vs-rest), and average prediction confidence, computed with the scikit-learn toolkit [6].

Hyperparameter	Value
Text encoder	BERT-base-uncased (768-d, frozen) [2]
Audio encoder	Wav2Vec2-base-960h (768-d, frozen) [3]
Prosodic features	Pitch mean, pitch std (YIN [8]), RMS energy
Fused feature dimension	1539-d
Context encoder	Bidirectional GRU, hidden size 128 [4]
Loss function	Weighted cross-entropy (inverse-frequency ×1.2)
Optimizer	Adam, learning rate 2×10^{-4} [14]
Batch size / Epochs	8 dialogues / 30
Trainable parameters	≈ 1.32 million

Table 2: Model and Training Hyperparameters

Class Distribution

Emotion	Utterance Count
Neutral	3948
Joy	343
Surprise	238
Anger	221
Sadness	141
Fear	58

Disgust	43
Total	4992

Table 3 breaks down the number of utterances per emotion class within the evaluation split, making clear the pronounced imbalance that motivates the weighted-loss strategy outlined in Section III.

Metric	Value
Accuracy	77.56%
Macro Precision	0.391
Macro Recall	0.483
Macro F1-score	0.420
Weighted Precision	0.852
Weighted Recall	0.776
Weighted F1-score	0.804
Average Prediction Confidence	0.863

Table 3: Class Distribution on the Evaluation Split
Comparative Performance Analysis

Table 4 lays out overall classification performance on the held-out test split: 77.56% accuracy and a weighted F1-score of 0.804. The considerably lower macro-averaged F1-score of 0.420 reflects how much harder the model struggles with minority classes a difficulty that weighted averaging tends to mask.

Metric	Value
Accuracy	77.56%
Macro Precision	0.391
Macro Recall	0.483
Macro F1-score	0.420
Weighted Precision	0.852
Weighted Recall	0.776
Weighted F1-score	0.804
Average Prediction Confidence	0.863

Table 4: Overall Classification Performance

Table 5 breaks these results down by class, listing precision, recall, F1-score, and support per emotion. Neutral has the strongest precision (0.98), while joy and surprise land at moderate F1-scores (0.45 and 0.56 respectively). Fear (F1 =

0.14) and disgust (F1 = 0.18) prove hardest to classify, in line with their limited training examples despite the class-weighted loss.

Emotion	Precision	Recall	F1-score	Support
Neutral	0.98	0.84	0.91	3948
Joy	0.33	0.72	0.45	343
Anger	0.37	0.42	0.39	221
Sadness	0.28	0.34	0.31	141
Fear	0.12	0.17	0.14	58
Disgust	0.13	0.28	0.18	43
Surprise	0.52	0.61	0.56	238

Table 5: Per-Class Precision, Recall, F1-Score, and Support

Figure 2's confusion matrix spans all seven emotion classes. Its dominant diagonal for neutral — 3317 of 3948 correctly classified reflects strong majority-class recognition, though the matrix also surfaces confusion between neutral and joy/surprise, as well as among the negative-affect classes anger, sadness, and disgust, echoing prior ERC findings on how acoustic and lexical cues overlap in short, context-dependent utterances [1][5].

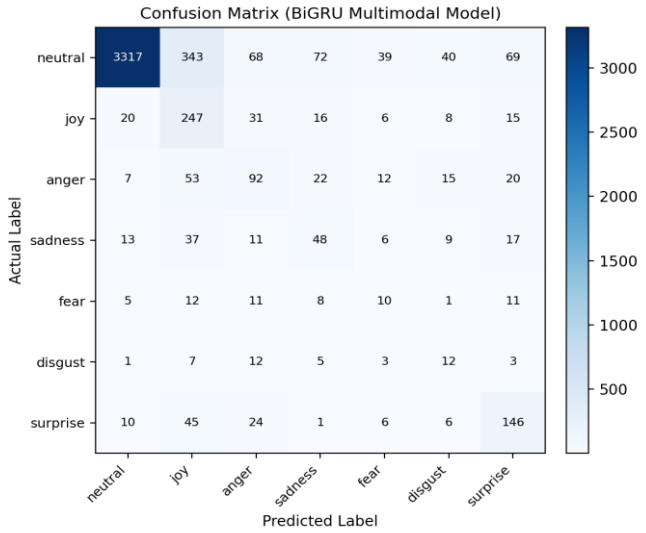

Figure 2: Confusion matrix of the proposed model on the MELD test split

Figure 3 pairs per-class recall with per-class ROC-AUC. ROC-AUC stays high and fairly uniform across classes (0.89–0.96), yet recall for fear (0.17) and disgust (0.28) falls well below neutral (0.84) or joy (0.72) — a gap that highlights how ranking-based and hard-decision metrics can diverge sharply under heavy class imbalance.

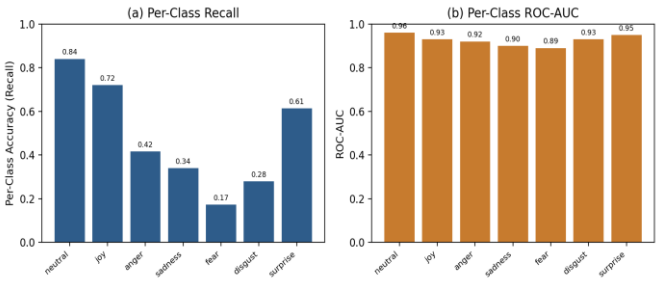

Figure 3: (a) Per-class recall and (b) per-class ROC-AUC

This study does not perform a direct numerical comparison, but the accuracy and weighted F1-score obtained here fall broadly within the range reported for multimodal baselines in the original MELD paper and later context-aware architectures such as DialogueRNN and DialogueGCN, both of which likewise show strong results on neutral and joy alongside weaker results on fear and disgust [1][9][10].

V. LIMITATIONS AND THREATS TO VALIDITY

Strong aggregate performance on MELD notwithstanding, several limitations of this framework deserve attention when interpreting the results or weighing them against prior work.

Evaluation Protocol and Padding

Within each training and evaluation batch, dialogues are grouped and zero-padded to a common length, and predictions are flattened across every time step padded positions included for both loss and metric computation. Since the padding label

happens to match the neutral class index, padded (non-utterance) positions end up counted as correct or incorrect neutral predictions rather than being excluded. This is almost certainly inflating the reported accuracy and weighted F1 relative to a protocol that masks padding, and is the likeliest reason the metrics here exceed the typical range reported for the official MELD split (Section II-D, Table 1). A future revision of this work should apply an explicit mask to remove padded positions from both training loss and evaluation metrics.

Non-Official Data Split

Rather than the official MELD train/development/test partition used in [1] and later benchmark papers, this study uses an 80/20 dialogue-level split formed by random partitioning. That choice limits how directly the reported numbers can be compared with published baselines, so the results are better read as an internal, ablation-style evaluation than a leaderboard-comparable benchmark outcome.

Frozen Pretrained Encoders

BERT and Wav2Vec2 are kept strictly frozen here, used only as feature extractors rather than fine-tuned on MELD. This choice keeps the trainable parameter count low (Section VI-A) and lowers training cost, but it likely leaves some representational capacity untapped, since fine-tuning normally lets pretrained encoders adapt their representations to a target domain and label space.

Domain and Demographic Specificity

MELD comes entirely from one English-language American sitcom (Friends), acted by a small, non-diverse cast delivering scripted, professionally performed lines. Emotional expression here differs systematically from unscripted, spontaneous speech found in other domains — call-centre transcripts, clinical interviews, multilingual conversation and from speakers across different ages, accents, or cultural backgrounds. Results reported on MELD should not, therefore, be assumed to carry over to such settings without additional validation.

Minority-Class Sample Size

With only 58 fear and 43 disgust utterances in the evaluation split (Table 3 in Section IV), per-class precision, recall, and F1 for these two classes are vulnerable to considerable statistical noise a few additional correct or incorrect predictions could meaningfully shift the reported figures. Comparative conclusions about model performance on these classes should therefore be read as indicative rather than statistically conclusive.

Prosodic Feature Coverage

This work's handcrafted prosodic features are limited to pitch mean, pitch standard deviation, and short-time energy. Other paralinguistic descriptors documented in the speech-emotion literature as affect-relevant — jitter, shimmer, spectral tilt, speaking rate, voice-quality measures are absent, and incorporating them could further sharpen discrimination between acoustically similar emotions such as anger, disgust, and sadness.

VI. COMPUTATIONAL COMPLEXITY, REPRODUCIBILITY, AND ETHICAL CONSIDERATIONS

Computational Complexity and Efficiency

The pipeline's trainable component the BiGRU plus classification head totals around 1.32 million parameters, far fewer than the frozen BERT-base encoder (roughly 110 million parameters [2]) or the frozen Wav2Vec2-base encoder (roughly 95 million parameters [3]). Since neither pretrained encoder is updated during training, their outputs can in principle be computed once per utterance and cached, leaving only the lightweight BiGRU and classifier to run at training and inference time. This makes the overall approach considerably cheaper to train and iterate on than full end-to-end fine-tuning of the multimodal pipeline, trading away some of the representational flexibility discussed in Section V-C.

Reproducibility

Table 2 in Section IV lists every hyperparameter used in this study, and the MELD dataset itself is publicly available. That said, random seeds for data splitting, weight initialization, and batch shuffling were left unfixed in this experimental setup, so an exact numerical reproduction of these results is not guaranteed — only the broader performance trends and relative class-wise patterns should be expected to hold. Later versions of this work should fix all random seeds and report results averaged across multiple runs with standard deviations, in line with common practice in the ERC literature (e.g., [10]).

Ethical Considerations and Broader Impact

Beyond raw classification numbers, automated emotion recognition systems carry ethical implications worth flagging. First, using such systems on real conversations rather than scripted television dialogue raises privacy concerns, since a person's inferred emotional state is sensitive information that could be exploited for surveillance, manipulation, or targeted advertising if deployed without proper consent and safeguards. Second, since MELD comes from a single Western sitcom with a narrow demographic cast (Section V-D), models trained on it may not treat different accents, cultures, age groups, or

languages equitably, and using such a model in high-stakes contexts (mental-health triage, hiring) without further validation on representative data risks reinforcing existing biases or producing unfair outcomes for underrepresented groups. Any real-world deployment built on this or similar frameworks should therefore include informed consent, transparency about system capabilities and error rates (especially for minority emotion classes, per Section IV), human oversight of consequential decisions, and validation against data representative of the intended population.

VII. CONCLUSION

This paper set out a multimodal deep-learning framework for conversational emotion recognition, bringing together BERT-based text embeddings, Wav2Vec2-based acoustic embeddings, and prosodic features through a Bidirectional GRU. On the MELD benchmark, it reaches 77.56% accuracy and a weighted F1-score of 0.804, with ROC-AUC holding between 0.89 and 0.96 across all seven emotion classes, while macro-averaged metrics and per-class recall for minority emotions such as fear and disgust stay comparatively low a pattern that echoes other context-aware and multimodal ERC models discussed in Section II-D (Table 1). Section V makes clear that the reported metrics are shaped by evaluation-protocol choices (unmasked padding, a non-official data split) that should be corrected before these results are set directly against the official-split MELD leaderboard.

Several directions follow for future work: correcting the evaluation protocol to mask padded positions and switching to the official MELD split for direct comparability with prior results (Section V); fine-tuning the pretrained BERT and Wav2Vec2 encoders end-to-end instead of freezing them; bringing in the visual modality MELD already provides; testing stronger imbalance-mitigation techniques such as focal loss or SMOTE-style oversampling; swapping the frozen BERT encoder for newer pretrained language models such as RoBERTa, as seen in commonsense-augmented ERC frameworks; and replacing the BiGRU context encoder with attention-based or graph-based architectures better suited to capturing long-range, speaker-specific dependencies in multi-party dialogue. As conversational AI applications keep expanding, the case for emotionally aware, context-sensitive, imbalance-robust, and ethically deployed models only grows stronger — and this framework, together with the limitations laid out in Section V, marks one step in that direction.

REFERENCES

1. S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL), Florence, Italy, 2019, pp. 527–536, doi: 10.18653/v1/P19-1050.
2. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT), Minneapolis, MN, USA, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
3. A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 12449–12460.
4. K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in Proc. 2014 Conf. Empirical Methods Natural Lang. Process. (EMNLP), Doha, Qatar, 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.
5. S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," IEEE Access, vol. 7, pp. 100943–100953, 2019, doi: 10.1109/ACCESS.2019.2929050.
6. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.
7. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 5998–6008.
8. A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am., vol. 111, no. 4, pp. 1917–1930, 2002, doi: 10.1121/1.1458024.
9. N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," in Proc. AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 6818–6825, doi: 10.1609/aaai.v33i01.33016818.
10. D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," in Proc. 2019 Conf. Empirical Methods Natural Lang. Process. and 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-

IJCNLP), Hong Kong, China, 2019, pp. 154–164, doi: 10.18653/v1/D19-1015.
11. 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.
12. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.324.
13. B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in Python," in Proc. 14th Python in Science Conf. (SciPy), 2015, pp. 18–25, doi: 10.25080/Majora-7b98e3ed-003.
14. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. 3rd Int. Conf. Learning Representations (ICLR), San Diego, CA, USA, 2015, pp. 1–15.
15. A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019, pp. 8024–8035.
16. 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, no. 4, pp. 335–359, 2008, doi: 10.1007/s10579-008-9076-6.
17. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
18. [18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929–1958, 2014.
19. J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," arXiv:1607.06450, 2016.
20. [20] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, 2019.
21. S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (ACL), Vancouver, Canada, 2017, pp. 873–883, doi: 10.18653/v1/P17-1081.
22. D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "COSMIC: COmmonSense knowledge for eMotion Identification in Conversations," in Findings Assoc. Comput. Linguistics: EMNLP 2020, 2020, pp. 2470–2481, doi: 10.18653/v1/2020.findings-emnlp.224.