Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 10, No. 2, April, 2026

Multimodal Emotion Recognition Using Hybrid Large Language Models and Metaheuristic Algorithms

Andino Maseleno [1, 2], M. Teduh Uliniansyah [2], Agung Santosa [2], Lyla Ruslana Aini [2], Rini Wijayanti [2], Ahmad Fudholi [3], Chotirat Ann Ratanamahatana [1*]

[1] *Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand.*

[2] *Research Organization for Electronics and Informatics, National Research and Innovation Agency, Jakarta Pusat 10340, Indonesia.*

[3] *Pusat Pengajian Citra Universiti, Universiti Kebangsaan Malaysia, Bangi, Selangor 43600, Malaysia.*

Abstract

Emotion recognition is a vital component of human–computer interaction and intelligent systems, yet robust multimodal emotion recognition remains challenging due to high-dimensional input space, noisy features, and the complexity of integrating heterogeneous modalities. This study proposes a novel hybrid multimodal framework that enhances both accuracy and computational efficiency by combining the semantic representation capability of Large Language Models (LLMs) with the optimization strengths of metaheuristic algorithms. In the proposed approach, an LLM is utilized to extract high-level contextual features from text and audio streams, while the Binary Artificial Hummingbird Algorithm (BAHA) performs feature selection to remove redundant attributes. Subsequently, the Goose Algorithm (GA) optimizes classifier hyperparameters, and the Komodo Mlipir Algorithm (KMA) conducts late fusion of the final multimodal outputs. Experiments conducted on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset, evaluated on six emotion categories, demonstrate that this hybrid approach successfully captures subtle affective cues and surpasses state-of-the-art baselines, achieving an accuracy of 87.5%. Integrating LLMs with multiple specialized metaheuristics therefore yields a substantially more robust emotion recognition pipeline and represents a promising direction toward the development of more emotionally intelligent systems.

Keywords:

Multimodal Emotion Recognition;
Large Language Models;
Metaheuristic Algorithms;
Binary Artificial Hummingbird Algorithm;
Goose Algorithm;
Komodo Mlipir Algorithm.

Article History:

Received:	06	October	2025
Revised:	22	February	2026
Accepted:	03	March	2026
Published:	01	April	2026

1- Introduction

Emotion recognition has emerged as a crucial component in intelligent systems, enabling more natural and adaptive interactions between humans and machines [1]. It has broad applications in domains such as mental health monitoring, education, customer service, and affective computing [2]. While unimodal emotion recognition methods (e.g., using only speech or text) have shown promising results, they often fail to capture the full spectrum of emotional cues [3]. Human emotions are inherently multimodal, conveyed simultaneously through linguistic content, vocal tone, facial expressions, and other behavioral signals [4]. Consequently, multimodal emotion recognition (MER) is increasingly recognized as a more comprehensive and reliable approach to understanding affective states [5]. Despite the potential of MER, several open challenges persist [6]. First, the high dimensionality of multimodal data can introduce redundant and noisy features that degrade model performance. Second, the heterogeneity of different modalities makes it difficult to develop unified feature representations and alignment strategies. Third, effective fusion techniques are required to combine information from multiple modalities without losing contextual meaning. Finally, optimizing deep learning

* **CONTACT**: chotirat.r@chula.ac.th

DOI: https://doi.org/10.28991/ESJ-2026-010-02-015

© 2026 by the authors. Licensee ESJ, Italy. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC-BY) license (https://creativecommons.org/licenses/by/4.0/).

models for emotion recognition is computationally intensive [7], and manual hyperparameter tuning often yields suboptimal results. These challenges motivate the need for advanced solutions in feature selection, model optimization, and multimodal fusion for MER. Recent advances in large language models (LLMs) offer new opportunities to address some of these issues [8]. LLMs excel at extracting high-level semantic and contextual information, particularly from text and transcribed speech—the key modalities in many MER scenarios. For example, an LLM can interpret the nuanced meaning of a sentence like "*I can't believe this happened again*" to determine whether it expresses sadness, frustration, or anger, based on linguistic context. However, LLMs alone are not sufficient to handle feature redundancy, parameter optimization, and multimodal fusion effectively [9].

To overcome these challenges, this research proposes a hybrid framework that integrates LLMs with metaheuristic algorithms for enhanced multimodal emotion recognition. Specifically, the BAHA is employed for feature selection to retain only the most discriminative attributes. The GA is adopted for hyperparameter tuning, ensuring optimal performance of the classification models. Furthermore, the KMA is introduced as a late fusion mechanism to combine outputs from different modalities into a robust final decision. By leveraging these complementary techniques, our approach aims to achieve higher accuracy, robustness, and computational efficiency. Recently, metaheuristic algorithms have demonstrated strong capabilities in emotion recognition.

Pan et al. [10] used the Genetic Algorithm to search for the most discriminative features from the geometric features of facial expressions and acoustic features of speech, by encoding feature subsets as binary strings and evolving these solutions to maximize recognition accuracy while minimizing model complexity. Chaudhari et al. [11], the metaheuristic algorithm in GCM-Net is primarily used for harmonic optimization of the feature set. Specifically, it assists in selecting the most relevant features for emotion and sentiment analysis by exploring the feature space extensively. Parameters such as the number of agents (set to 4) and the number of iterations (set to 100) are tuned to leverage the local search capabilities of the algorithm, ensuring an optimal subset of features is chosen. This optimized feature selection process enhances the model's ability to effectively fuse multimodal data and improves overall performance. Additionally, the algorithm's parameters, such as alpha and gamma (values of 0.6 and 0.5 respectively), are tuned considering dataset imbalance, further refining the classifier's effectiveness.

Michael et al. [12] focused on improving how well computers can recognize emotions in speech by tuning the parameters of a specific model, Long Short-Term Memory (LSTM), used in speech emotion recognition. It compares regular manual tuning with a genetic algorithm approach. The metaheuristic algorithm, specifically the genetic algorithm, plays a central role in optimizing the hyperparameters of the LSTM-based speech emotion recognition model. The main role of the genetic algorithm is to automatically identify the best combination of hyperparameters, such as batch size, dropout rate, number of LSTM neurons, and learning rate, that maximize the model's performance metrics, such as accuracy, precision, recall, and F1 score.

Mukta et al. [13] focused on predicting General Self-Efficacy (GSE) from Facebook data by analyzing both profile pictures and statuses. It combines traditional tools like Linguistic Inquiry and Word Count (LIWC) and Mediapipe with advanced deep learning techniques such as BERT, Convolutional Neural Networks (CNNs), and image segmentation models. Two approaches — tool-based and deep learning-based — were used to extract features, followed by machine learning classifiers to predict GSE scores. The tool-based model achieved up to 93.25% accuracy when combining text and image features, while the deep learning model was slightly less accurate but still competitive. This approach helps understand a person's self-belief through social media data. The metaheuristic role in this research is primarily centered around the use of Particle Swarm Optimization (PSO). PSO is employed for feature selection and hyperparameter tuning within the modeling framework. Specifically, it helps identify the most relevant features from the combined tool-based features, optimizing the selection process to enhance model performance. Additionally, PSO's ability to efficiently navigate large hyperparameter spaces allows for better tuning of the predictive models, potentially leading to improved accuracy and convergence compared to traditional search methods. This application of PSO contributes to refining the feature set and optimizing the parameters of the machine learning models used for GSE prediction, thereby enhancing the overall efficacy of the proposed approach

Daneshfar et al. [14] presented a hybrid framework for recognizing speech emotions that incorporates feature extraction, dimensionality reduction, and classification. The metaheuristic role in this research is primarily embodied in the development and application of the modified quantum-behaved particle swarm optimization (QPSO) algorithm, referred to as pQPSO. This algorithm serves as a metaheuristic optimization technique used for the discriminative dimensionality reduction stage. Specifically, it searches for an optimal projection matrix that maximizes emotion classification accuracy by effectively exploring the feature space and avoiding local minima, which are common challenges in high-dimensional optimization problems. The adaptive contraction-expansion factor and the use of truncated Laplace distribution (TLD) for solution generation enhance its exploration and exploitation capabilities, thereby improving the overall performance of the emotion recognition system.

However, to the best of our knowledge, no prior work has unified the strengths of LLMs and metaheuristic optimization within a single MER framework. The novelty of our approach lies in this integration: our framework

simultaneously tackles feature redundancy (via BAHA-based selection), model hyperparameter tuning (via GA), and multimodal fusion (via KMA) on top of advanced LLM-driven feature representations. In doing so, we fill a significant gap in the literature by providing a holistic solution that addresses multiple facets of the MER problem in tandem.

In summary, the main contributions of this paper are as follows:

- We introduce a unified multimodal emotion recognition framework that combines an LLM with three metaheuristic algorithms (BAHA, GA, and KMA). This novel integration exploits the complementary strengths of deep semantic modelling and evolutionary optimization.
- The framework employs the LLM to extract high-level semantic features from text and audio modalities, improving the system's ability to capture nuanced emotional cues.
- BAHA is utilized for feature selection, which reduces data dimensionality and filters out noisy or irrelevant features, thereby enhancing efficiency and classification accuracy.
- GA is applied for hyperparameter tuning of the learning models, providing an automated optimization mechanism that yields better performance than manual tuning.
- KMA is implemented for late fusion of the modality-specific outputs, adaptively weighting and combining the contributions of each modality to produce a more reliable final emotion prediction.
- Extensive experiments on the IEMOCAP benchmark (covering six emotions) show that our hybrid approach achieves an accuracy of 87.5%, outperforming state-of-the-art methods. This result demonstrates the effectiveness of the proposed framework in improving emotion recognition performance across diverse emotional categories.

The remainder of the study is organized as follows. Section 2 explains the related works. In section 3, the proposed methodology details, large language models, metaheuristic optimization, binary artificial hummingbird algorithm, GOOSE algorithm, and Komodo mlipir algorithm are explained. Results and discussion are presented in section 4. Finally, conclusions are given in section 5.

2- Related Works

This section provides a comprehensive review of existing research on multimodal emotion recognition using large language models. Based on the existing literature, there are no other works for multimodal emotion detection using hybrid large language models, binary artificial hummingbird algorithm for feature selection, goose algorithm for hyperparameter tuning, and late fusion komodo mlipir algorithm. Dutta et al. [15] introduced an approach to pretrain a text-based emotion recognition model using unsupervised speech transcripts guided by LLMs. These transcripts are generated from raw speech data through a pre-trained ASR system, after which an LLM is prompted to assign pseudo-labels. The resulting pseudo-labelled transcripts are then used to train an utterance-level text-based emotion recognition model. Meanwhile, Hong et al. [16] developed zero-shot and few-shot prompting techniques that leverage previous dialogue as contextual information for handling ambiguous emotions. Their experiments across three datasets demonstrate the strong potential of LLMs in recognizing ambiguous emotions and emphasize the advantages of incorporating contextual cues. Kim et al. [17] proposed a more efficient modality fusion approach, called MMER-LMF, designed to address the scarcity of Korean emotion datasets and enhance emotion recognition accuracy while minimizing training complexity.

Li et. al. [18] introduced Task-Specific Feature Learning (TSFL) method with three modules: Two-Stream LLM Feature Extraction, Coarse-Grained Task-Specific Feature Decomposition, Fine-Grained Task-Specific Feature Learning. Lu et al. [19] proposed an Emotion-Action Interpreter based on a Large Language Model (EAI-LLM), which produces textual interpretations by representing 3D body motion data as distinctive input tokens within LLMs. Teng et al. [20] proposed a depression detection pipeline that generates emotion prompts tailored to individual data. This approach integrates cross-modality fusion via cross attention mechanisms to combine depressive and emotional features, creating a comprehensive representation of depression indicators.

Xu et al. [21] proposed Reliable Learning Framework (RLF) includes a Hierarchical Interaction Network and a Reliable Fusion Strategy. The former can excavate emotion and intent cues from the high-level semantic features of multimodal data (video, audio, and text) generated by pretrained Large Language Models (LMMs), to enhance their representations, and the latter reliably integrates multiple predictions to further improve the robustness of emotion and intent understanding. Yacoubi et al. [22] developed a model based on adaptive fine-tuning using LoRA on LLaMA-3-8B, which enabled to highlight the remarkable ability of LLMs to capture the subtle nuances of emotions within textual data. Yang et al. [23] proposed Multimodal Sentiment Analysis and Emotion Recognition Adapter (MSE-Adapter), a lightweight and adaptable plugin. This plugin enables a large language model (LLM) to carry out MSA or ERC tasks with minimal computational overhead (only introduces approximately 2.6M to 2.8M trainable parameters upon the 6/7B

models), while preserving the intrinsic capabilities of the LLM. Within the MSE-Adapter, the Text-Guide-Mixer (TGM) module is employed to create explicit links between textual and non-textual modalities using the Hadamard product. This mechanism enables non-textual modalities to more effectively align with textual ones at the feature level, thereby enhancing the production of higher-quality pseudo tokens.

Zhang et al. [24] initiated the first attempt by proposing IntervEEG-LLM, an end-to-end framework leveraging EEG-based multimodal signals for mental health intervention. IntervEEG-LLM first used LLMs for heterogeneous multimodal physiological understanding of signals and predicted the intermediate modeling results of users. Then it leveraged effective intervention strategies to interpret and apply the modeling results, offering concrete intervention text through natural language interfaces. Zhang et al. [25] modelled emotion recognition is formulated as a text generation task; therefore, DialogueLLM is introduced, a large language model fine-tuned with context and emotion knowledge to enhance performance. Zhang et al. [26] proposed the Text-guided Multimodal Emotion Intent Joint Recognition method. By leveraging the text modality to guide the fusion process, it reduces the noise introduced by other modalities. To strengthen the text modality's guiding role, used large language models (LLMs) for multiturn targeted data augmentation and oversampling strategies to address data imbalance. Zhou et al. [27] proposed to explore emotional distribution information in interviews to assist multi-modal ADD model. On one hand, used large language models (LLMs) to automatically recognize emotion of text data, and re-organize the data guided by the valence attribute of emotion, which facilitates their model being aware of difference in emotion distribution. On the other hand, designed the emotion encoding which enhances the proposed model to consider the emotional distribution information in its decision-making process. Table 1 shows related works.

Table 1. Related works

No.	Year	Author	Modality	Dataset	Method	Results
1	2025	Dutta et al. [15]	Audio, text	IEMOCAP, MELD, CMU-MOSI	Pre-training model with speech-to-text transcripts using Whisper-large model for emotion labeling, subsequently fine-tunes a RoBERTa model for emotion recognition from text, while employing a Bi-GRU combined with cross-attention fusion.	Achieved state-of-the-art results on IEMOCAP and CMU-MOSI datasets. Performance improvements include 16.46% (audio) and 20.06% (text) for IEMOCAP after hierarchical training stages. Multi-modal fusion achieved 16.28% improvement for IEMOCAP.
2	2025	Hong et al. [16]	Text	MSP-Podcast, IEMOCAP, and GoEmotions	The study employs LLMs, specifically Gemini-1.5-Flash, as the backbone for emotion recognition.	Accuracy: 56%
3	2025	Kim et al. [17]	Video, text	10,351 Korean video datasets taken from AI-HUB	• Leveraging large pre-trained language models to overcome the scarcity of Korean emotion datasets • Utilizing multimodal integration of text and video to achieve improved accuracy • Employing emotion score-based decision fusion to balance performance across modalities.	Text modality (GPT-4 Turbo): 58% accuracy for simple expressions, 72% accuracy with contextual information. Video modality (3D CNN):74% accuracy.
4	2025	Li et al. [18]	Video, audio, text	MC-EIU dataset (31,320 training samples, 4,474 validation samples)	Task-Specific Feature Learning (TSFL) method with three modules: • Two-Stream LLM Feature Extraction, • Coarse-Grained Task-Specific Feature Decomposition, • Fine-Grained Task-Specific Feature Learning.	Achieved a JRBM score of 0.6230, outperforming the official baseline and winning the MEIJU challenge.
5	2025	Lu et al. [19]	Visual	the Emilya dataset, the KDAE dataset, and the EGBM corpus	EAI-LLM produces textual interpretations by representing 3D body motion data as distinctive input tokens within LLMs. It employs a multi-granularity skeleton tokenizer specifically designed for LLMs, which independently extracts spatio-temporal tokens and semantic tokens from the skeletal data.	Accuracy: KDAE 71.17%, EGBM 66.97%
6	2025	Teng et al. [20]	Text, audio, video	E-DAIC, EATD	Generate emotion prompts from text using a pretrained LLM, extract features from text, video, audio using pretrained BERT, and prompts, fuse them via cross-attention (intra-modality) and inter-modality fusion, then use the combined features to predict depression levels.	CCC of 0.688 and an MAE of 3.89 on the E-DAIC dataset. On the EATD dataset, it achieves an F1 Score of 0.83, Precision of 0.79, and Recall of 0.87.
7	2025	Xu et al. [21]	Video, audio, text	MC-EIU (Mandarin subset) with 11,003 utterances (training, validation, and test sets)	Reliable Learning Framework (RLF) consisting of Hierarchical Interaction Network (LF-HIN) and Reliable Fusion Strategy (RFS), utilizing pretrained models for emotion and intent recognition.	Achieved 0.7285 for emotion, 0.7456 for intent, and 0.7370 for joint understanding, securing first place in the MEIJU Challenge Track 2 (Mandarin).

8	2025	Yacoubi et al. [22]	Text	ISEAR, Emotion for NLP and SemEval 2019	• Adaptive Fine-Tuning Analysis – A thorough examination of adaptive fine-tuning methods and their effects on emotion recognition performance. • Novel LoRA-Based Approach – Introduction of a new text-based emotion detection method by fine-tuning Llama-3-8B with LoRA, tailoring it for emotion recognition tasks.	ISEAR F1-score of 79%, Sem-Eval F1-score of 92%, NLP dataset F1-score of 88%
9	2025	Yang et al. [23]	Text, Audio, Vision	MOSEI, CH -SIMS-V2, MELD, CHERMA	Multimodal Sentiment Analysis and Emotion Recognition Adapter (MSE-Adapter) plugin with modules: Text-Guide-Mixer (TGM) for feature alignment and Multi-Scale Fusion (MSF) for early fusion of non-textual modalities.	MSE-Adapter achieved competitive performance across datasets, with MSE-ChatGLM3-6B outperforming baselines, especially on the CHERMA dataset. The plugin showed minimal computational overhead (2.6M-2.8M trainable parameters), preserving LLM's general capabilities while boosting performance.
10	2025	Zhang et al. [24]	Text	MELD, IEMOCAP, EmoryNLP	DialogueLLM, a large language model specifically fine-tuned with conversational and emotional knowledge to enhance emotion recognition in dialogues.	MELD: 71.91%; IEMOCAP: 70.48%; EmoryNLP: 41.76%
11	2025	Zhang et al. [25]	EEG, text, and dialogue data	EEG-based mental health database (MODMA and crafted appropriate inputs according to AMIGOS database.	IntervEEG-LLM is a two-stage framework that: • Predicts mental states from EEG and dialogue data. • Generates natural language interventions by combining dialogue context, predictions, and evidence-based strategies (physical activity, mindfulness, meditation, mood tracking, gratitude). • It uses a four-layer architecture: data processing, LLM modeling, LLM reasoning, and LLM dialogue, to integrate multimodal inputs, apply mental health strategies, and deliver empathetic, real-time responses.	IntervEEG-LLM effectively adapts EEG-based multimodal data to deliver context-specific, evidence-based mental health interventions through personalized, transparent dialogue
12	2025	Zhang et al. [26]	Text, audio, video	Emotion and Intent Joint Understanding in Multimodal Conversation (MC-EIU)	Trains a DeBERTaV3 text classifier, augmented with a three-round LLM-based data augmentation strategy to address data imbalance. multimodal classification → extracts features from text (DeBERTaV3), video (OpenFace + ResNet-50), and audio (Wav2Vec), fuses them via cross-attention	Emotion F1 score 0.4025, intent F1 score 0.4982, Online score 0.4641
13	2025	Zhou et al. [27]	Text, audio	E-DAIC, MODMA	Prompts an LLM to label emotions in interview text, reorders sentences by valence, encodes the emotional distribution, and fuses these embeddings with audio- text features via a pre-trained, attention-based multimodal depression detection model.	E-DAIC: ACC: 0.80, Precision: 0.64, Recall: 0.82, F1-score: 0.72, MAE: 4.05, RMSE: 5.00, CCC: 0.60. MODMA: ACC: 0.87, Precision: 0.82, Recall: 0.96, F1-score: 0.87, MAE: 6.14, RMSE: 7.44, CCC: 0.50.

3- Research Methodology

Metaheuristic algorithms are inspired by a wide range of natural and artificial phenomena, such as animal behavior, biological processes, genetics, physics, human activities, game theory, and evolutionary principles. Based on their sources of inspiration and design principles, these algorithms can be grouped into several categories:

1) *Swarm-based algorithms:* These methods replicate the collective behavior of social creatures, particularly their strategies for locating and exploiting food sources. 2) *Human-inspired metaheuristics:* Drawing on human actions and decision-making processes in different contexts, these algorithms attempt to mimic how humans' reason and choose optimal solutions to optimization problems. 3) *Evolutionary algorithms (EAs):* Among the earliest metaheuristics, these approaches are rooted in the analogy between biological evolution and optimization. Decision variables are treated as "genes," and processes such as crossover, mutation, and selection are applied to preserve and improve adaptive solutions across generations. 4) *Game-theory-based metaheuristics:* These algorithms employ principles of game theory, leveraging competitive and cooperative strategies to search for optimal solutions through strategic interactions. 5) *Physics-inspired metaheuristics:* Grounded in the laws of physics, these methods simulate phenomena such as gravitation, magnetism, and thermodynamics. By modeling interactions among search agents according to these physical processes, they guide the exploration and exploitation of the solution space to identify high-quality solutions for complex problems. Each category harnesses distinct principles—whether biological, social, strategic, or physical—to shape the search process, ultimately aiming to solve optimization problems effectively.

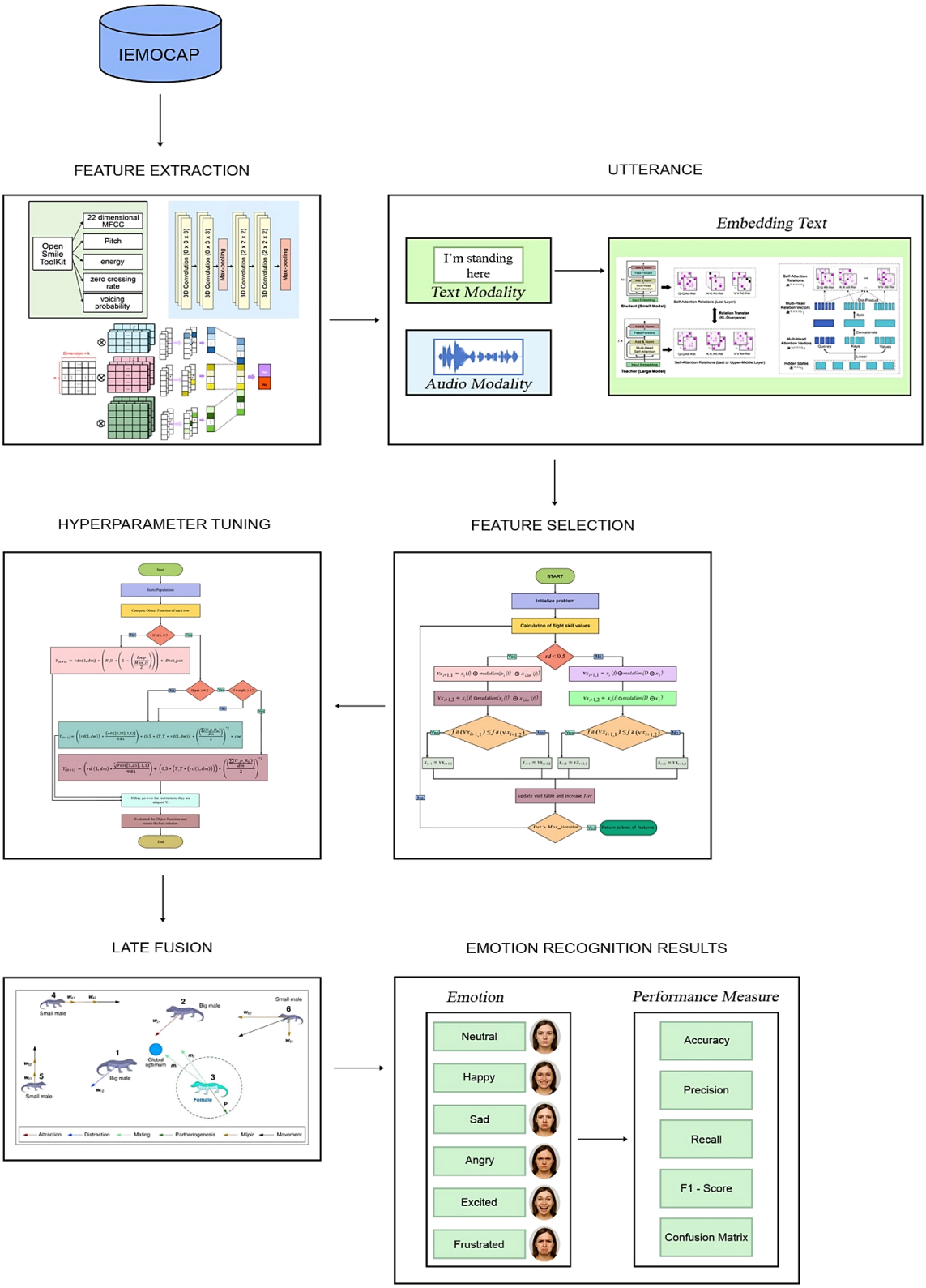

Figure 1. Proposed methodology

Figure 1 presents a complete framework for multimodal emotion recognition, beginning with the dataset and ending with the evaluation of classification results. The process starts with the dataset, which contains both text and audio modalities as the primary input sources for emotion analysis. These modalities are then passed to the feature extraction stage, where embeddings are generated for textual data and acoustic features such as prosody, spectral, and pitch-based attributes are derived from the audio signal. These extracted features form meaningful representations of the utterances that serve as input to subsequent stages. Once features are obtained, they undergo feature selection to retain only the most relevant information. This step is crucial for reducing dimensionality, eliminating noise, and ensuring that the retained features contribute to more effective classification. In parallel, hyperparameter tuning is carried out to optimize

the parameters of the learning models, improving their accuracy, stability, and generalization ability. These two processes ensure that the system works with high-quality input representations and optimally configured models. Following this, the framework employs a late fusion strategy to combine the outputs from the text and audio modalities. Unlike early fusion, which merges raw features, late fusion integrates decisions at a higher level, enhancing robustness by leveraging complementary strengths of each modality. The fused results are then fed into the final classification module, which identifies the target emotions.

This choice of late fusion over intermediate or hybrid approaches was mainly driven by two factors: the heterogeneity between text and audio representations, and the need to maintain modality independence during feature learning. This independence allows every modality to learn its own distinctive patterns without requiring strict alignment, which is helpful when their data distributions differ. As highlighted by Lian et al. [28], decision-level (late) fusion is simpler, more flexible, and does not require temporal synchronization across modalities, while allowing each modality to employ its own optimal classifier to enhance local decision quality. Empirical evidence from [29] further supports this choice, showing that a late fusion–based framework outperforms previous multimodal fusion methods on the IEMOCAP dataset. However, because late fusion assumes that each modality works independently, it may overlook useful interactions between modalities that could improve emotion recognition accuracy. Consequently, the trade-off lies between robustness and inter-modality synergy: late fusion offers better robustness and modularity, while intermediate or hybrid fusion can capture stronger cross-modal relationships but often with greater model complexity and sensitivity to data imbalance. The last stage of the framework presents the emotion recognition results, classifying utterances into six emotion categories: Neutral, Happy, Sad, Angry, Excited, and Frustrated. Alongside the predictions, system performance is assessed using common evaluation metrics, including accuracy, precision, recall, F1-score, and the confusion matrix. Together, these measures provide a comprehensive evaluation of how effectively the system recognizes emotions across modalities.

The selection of three distinct metaheuristics (BAHA, GOOSE, and KMA) for this framework was not an empirical choice but was theoretically grounded in the "No Free Lunch" (NFL) theorem [30]. The NFL theorem formally states that no single optimization algorithm can achieve superior performance across all classes of optimization problems [30, 31]. The efficacy of any heuristic is intrinsically tied to its adaptation to the specific topological structure of the problem's search space. This theorem provides the foundational justification for developing hybrid or modular frameworks that apply specialized solvers to distinct sub-problems, rather than attempting to use a single, universal algorithm.

Specifically, the framework decomposes the optimization challenge to balance the critical trade-off between exploration and exploitation. First, the feature selection task addressed by BAHA (Section 3.3) is a binary combinatorial optimization problem [32], and BAHA was selected for its native design to operate within this discrete search space. Second, the remaining tasks are decoupled: the GOOSE algorithm, which simulates the broad search behavior of a flock, is employed for the high-dimensional, global exploration required for hyperparameter tuning. This prevents premature convergence in the vast, non-convex hyperparameter landscape. In contrast, the Komodo Mlipir Algorithm (KMA), with its 'mlipir' movement, Equations 30 and 31, provides the strong local exploitation and convergence pressure necessary for the low-dimensional, continuous optimization task of fine-tuning the late-fusion weights. This separation of concerns—using distinct algorithms for discrete selection, global exploration, and local exploitation—is a theoretically-sound strategy to achieve robust and efficient optimization [30, 31].

3-1- Dataset

The experiments were performed using the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [33], which serves as a widely recognized benchmark for multimodal emotion recognition. Developed by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC), the IEMOCAP corpus contains around 12 hours of dyadic conversations recorded from ten professional actors. During these sessions, facial, head, and hand markers were used to capture detailed motion data, enabling fine-grained analysis of expressions and gestures in both scripted dialogues and spontaneous conversations. The actors performed selected emotional scripts as well as improvised scenarios designed to elicit specific affective states such as happiness, anger, sadness, frustration, and neutrality. Speaker segmented each dialogue turns, and the resulting conversational units were annotated by three independent raters for perceived emotional content. The annotations include both categorical emotion labels and continuous ratings along the valence, dominance, and activation dimensions. With its multimodal motion-capture detail, naturalistic and controlled elicitation settings, and comprehensive annotation scheme, IEMOCAP provides a rich and valuable resource for advancing research in multimodal emotion recognition and expressive human communication [34].

In this study, six categorical emotions were used for classification. Table 2 shows description of emotion. For each utterance, three modalities were considered: Audio (low-level descriptors): frame-level acoustic features (e.g., MFCCs, pitch, energy). Audio (high-level descriptors): higher-level prosodic and spectral features. Text (LLM embeddings): contextual embeddings obtained from transcripts using the sentence-transformers/all-MiniLM-L6-v2 model (384 dimensions). Figure 2 shows sample utterances in a text. The dataset was split into training (80%) and testing (20%) partitions, stratified across classes. A subset of the test set was further split into validation and holdout partitions to optimize late-fusion weights with KMA.

Table 2. Description of emotion

Emotion	Description
Neutral	A balanced emotional state in which a person does not feel strong positive or negative emotions.
Happy	A positive emotional state characterized by feelings of joy, contentment, and overall well-being.
Sad	A negative emotional state associated with feelings of loss, disappointment, or helplessness.
Angry	A strong negative emotional state that occurs when someone perceives unfair treatment, obstacles, or threats.
Excited	A strong, positive emotional state of heightened energy, enthusiasm, or eagerness.
Frustrated	A strong feeling of disappointment, annoyance, or dissatisfaction that arises when someone's goals, desires, or expectations are blocked or hindered.

Figure 2. Sample utterances in a text

3-2- Large Language Models

In recent years, major progress in natural language processing (NLP) has been driven by the rise of large language models (LLMs) [35]. These models demonstrate impressive abilities, including in-context learning, few-shot prompting, and instruction following, among others. These dynamic abilities have greatly contributed to boosting the effectiveness of language models, thus enabling AI algorithms to achieve unparalleled levels of effectiveness and productivity. Generative AI is a type of artificial intelligence capable of producing videos, texts, and images in response to prompts [36]. Generative AI is an innovative method that not only analyzes existing data but also creates entirely new content. These models learn patterns from vast datasets to generate outputs. The modern era of generative large language models (LLMs) began with the research paper "Attention is All You Need," in which the transformer architecture was first introduced [37]. Transformers are a framework that leverages the attention mechanism, enabling models to capture global dependencies within data. Figure 3 shows transformer architecture. Typically, models like the transformer architecture-based LLMs are first pre-trained using extensive datasets comprising diverse languages and domains. Shortly afterward, BERT (Bidirectional Encoder Representations from Transformers) emerged as a state-of-the-art model, offering a bidirectional and unsupervised approach to language representation.

The development of transformer-based models continued with GPT (Generative Pre-trained Transformers), followed by GPT-2, which demonstrated improved performance across various tasks without task-specific training, and GPT-3, which further scaled capabilities, particularly excelling at few-shot learning [38]. OpenAI has made notable contributions with the development of two landmark models, ChatGPT and GPT-4, which represent a transformative step in language processing. Nevertheless, because these models are proprietary, numerous alternative LLMs have emerged, many containing tens or even hundreds of billions of parameters. In this work, we seek to classify LLMs into two main categories based on their scope: general-purpose LLMs and specialized LLMs. General-purpose LLMs are built for broad applicability across diverse NLP tasks, such as machine translation, text comprehension, and dialogue generation. Well-known examples include GPT-4, Claude, ChatGPT, LLaMA, PanGu-Σ, Bard, and Falcon [39]. These models are not tailored for any single task; while they demonstrate strong performance across multiple domains, their effectiveness in highly specific applications remains an area for further investigation. In contrast, specialized LLMs—often referred to as task-specific LLMs—are adapted through fine-tuning with task-oriented architectures and domain knowledge. This enables them to deliver comparable or even superior performance to general LLMs, often with significantly fewer parameters. For instance, Chen et al. [40] introduced Phoenix, a large language model designed to support multiple languages. Similarly, Liu et al. [41] fine-tuned the Goat model, derived from LLaMA, to address arithmetic-related tasks.

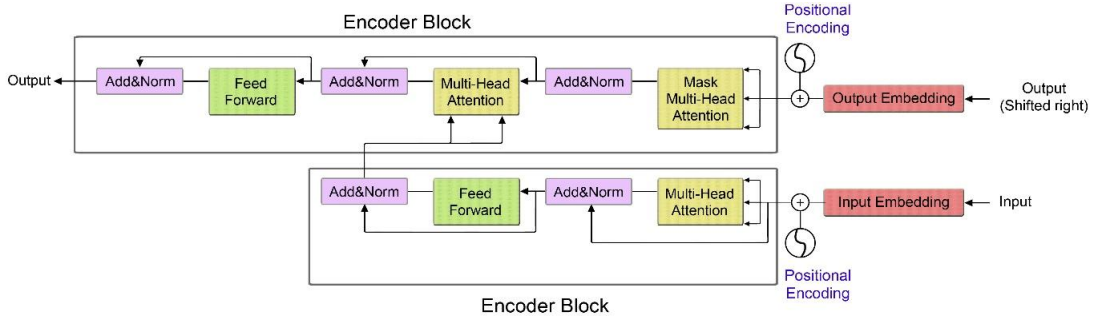

Figure 3. Transformer Architecture

In essence, the Transformer architecture is composed of multiple encoder and decoder layers. Each layer incorporates a specialized attention mechanism that allows the model to capture the contextual meaning of every word (or token) within a sentence. For example, the word *"park"* is represented differently in the phrases *"I walk my dog in the park"* and *"I park my car in the garage."* The encoder component of Transformers is mainly applied to classification-oriented tasks, such as sentiment analysis, question answering, and named entity recognition. Prominent models built on Transformer encoders include BERT, DistilBERT, and XLM-Roberta. In contrast, the decoder component is primarily employed for sequence generation tasks. A notable example is the GPT family, which forms the backbone of ChatGPT. Sentence Transformers utilize the encoder part of Transformers to generate the embedding of a sentence. Figure 4 shows sentence transformer.

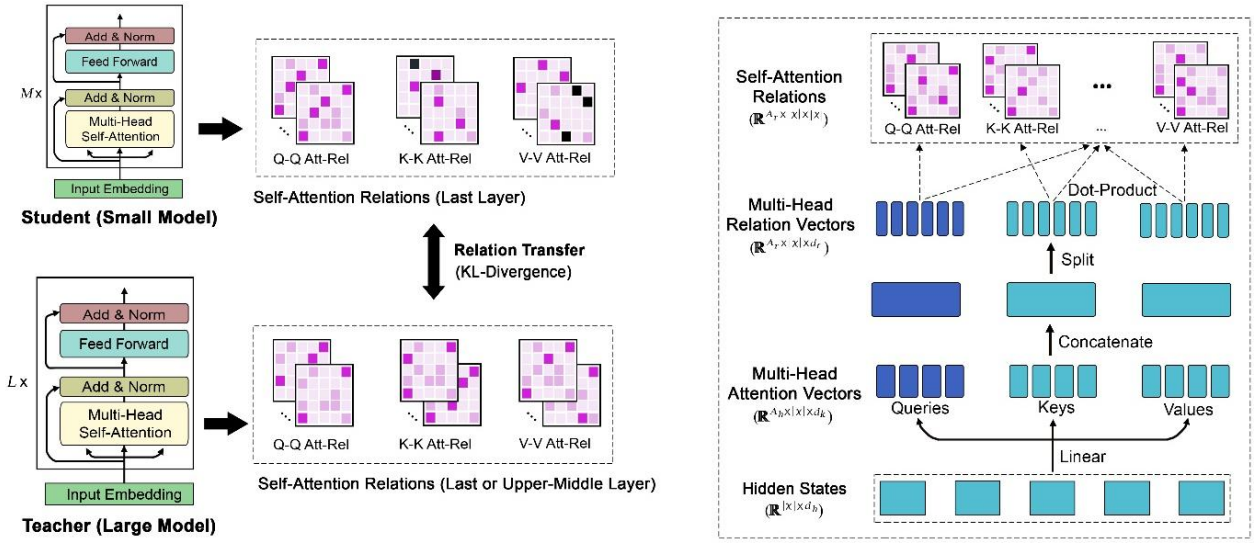

Figure 4. Sentence transformer

3-3- Binary Artificial Hummingbird Algorithm

The AHA algorithm operates continuously within its defined domain, meaning that the optimal positions it identifies are not restricted to binary values of zero and one [42]. To adapt the algorithm for binary space, a transfer function is employed. In this binary adaptation, all values are constrained to {0,1}, resulting in binary outputs. The process of binarization enhances the performance of the AHA algorithm by converting its arithmetic-based operators into logical operators. While the visiting table, its initialization, and the flight skills remain the same in both AHA and BAHA, differences arise in the population initialization and foraging strategies. Specifically, in BAHA, hummingbirds are assigned binary positions, and the population is initialized according to the following Equation 1.

$$x_{i,j} \in \{0,1\}, \qquad x_{i,j} = rand\{0,1\}, \;\; i = 1,\dots,n \,, \;\; j = 1,\dots,d \tag{1}$$

where, $rand\{0,1\}$ is a randomly selected element from the set {0,1}. The parameters *n* and *j* denote the population size and the problem's dimensionality, respectively. x*i,j* refers to the *j*th position in the solution vector of the *i*th food source. Once the population is initialized, the algorithm advances using foraging strategies. Figure 5 shows flowchart of BAHA algorithm. The following subsections present the foraging strategies employed in the BAHA algorithm.

3-3-1- Guided Foraging

The initial foraging strategy adopted by hummingbirds is guided foraging. In this approach, the hummingbird focuses on exploiting resources within its local neighborhood. To simulate this behavior, the following Equations 2 and 3 are applied.

$$Vx_i^{(t+1,1)} = x_i^{(t)} \oplus mutation\left(x_i^{(t)} \oplus x_{i,tar}^{(t)}\right) \tag{2}$$

$$Vx_i^{(t+1,2)} = x_i^{(t)} \odot mutation\left(x_i^{(t)} \oplus x_{i,tar}^{(t)}\right) \tag{3}$$

where, $x_i^{(t)}$ and $x_{i,tar}$ denote the position of the *i*th food source and the position of the target food source at time *t*, respectively. $Vx_i(t+1)$ represents the provisional position of the *i*th food source at time $t+1$. The symbols $\oplus$ and $\odot$ correspond to the logical AND and OR operators, respectively. The mutation function introduces random modifications to certain bits of the binary representation of food source positions. If $x_i = x_{i,1}x_{i,2}\dots x_{i,d}$ indicates the position of the *i*th food source in d dimensions (where d is the number of original features in the feature selection problem), then $x_{i,j}$ denotes the *j*th place in the position of the *i*th food source. Based on these definitions, the mutation function can be expressed as Equation 4:

$$mutation(x_i) = [m_{i,1}m_{i,2}, \ldots, m_{i,d}] \quad (4)$$

where, d denotes the dimensionality of the feature selection problem, and $m_{i,j}$ indicates the mutated jth element of the position vector corresponding to the ith food source.

The computed values from Equations 2 and 3 are compared, and the subsequent position at time t+1 is determined using the following Equation 5.

$$x_i^{(t+1)} \begin{cases} Vx_{i+1,1} \; if \; fit(Vx_{i+1,1}) \leq fit(Vx_{i+1,2}) \\ Vx_{i+1,2} \quad if \; fit(Vx_{i+1,1}) > fit(Vx_{i+1,2}) \end{cases} \quad (5)$$

3-3-2- Territorial Foraging

The second feeding strategy employed by hummingbirds is territorial foraging. This strategy corresponds to the exploration phase and helps prevent the proposed algorithm from becoming trapped in local minima. The territorial foraging process is modeled using the following Equations 6 and 7.

$$Vx_i^{(t+1,1)} = x_i^{(t)} \oplus mutation\left(D \oplus x_i^{(t)}\right) \quad (6)$$

$$Vx_i^{(t+1,2)} = x_i^{(t)} \odot mutation\left(D \oplus x_i^{(t)}\right) \quad (7)$$

where, $x_i^{(t)}$ denotes the position of the ith food source at time t, while D is derived from the hummingbird's flying ability. The symbols $\oplus$ and $\odot$ correspond to the logical AND and OR operators, respectively. The values obtained from Equations 6 and 7 are then compared, and the subsequent position at time t+1 is determined using the Equation 5. A detailed description of the BAHA process can be found in Hamdipour et al. [42].

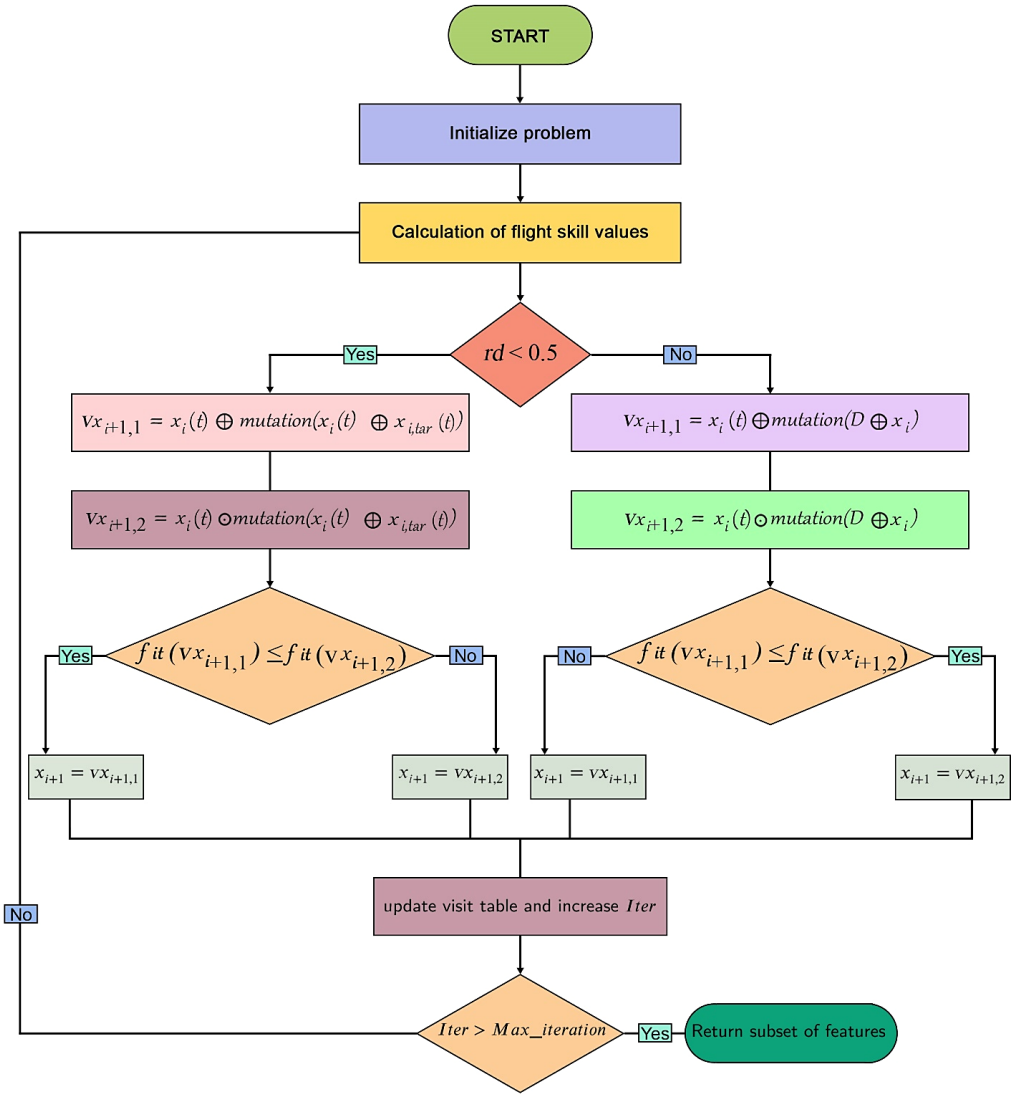

Figure 5. Flowchart of BAHA Algorithm

3-4- GOOSE Algorithm

GOOSE begins by initializing the population, referred to as the *Y* matrix. Each goose's position corresponds to an element within *Y*. After initialization, any search agents that move beyond the defined search space are returned to valid positions [43]. In every iteration, the fitness of each search agent is evaluated using standard benchmark functions. The fitness values of all agents (each row in the *Y* matrix) are then compared with those of the others to identify the *BestFitness* and the corresponding *Best Position* (*BestY*). *BestFitness* and *BestY* are operations that compare the fitness of each current row $(fitness_i + 1)$, and during iterations, the fitness of the row before it $(fitness_i)$ is given back.

In the subsequent steps, the balance between exploration and exploitation is regulated through a condition and a random variable. This variable is designed to distribute both phases evenly across the total number of iterations. In the GOOSE algorithm, a 50% probability is assigned to either exploration or exploitation, determined by a designated random variable, denoted as *rd*. Consequently, the iterations are evenly divided between the two phases using a conditional statement. Additionally, several other variables are introduced, including *pro*, *rd*, and *coe*, all of which take random values within the range [0,1]. A condition is applied to ensure that if *coe* is less than or equal to 0.17, its value is retained; otherwise, it is set to 0.17. The *pro* variable functions as a selector, determining which equation will be applied. Furthermore, the algorithm also incorporates a variable representing the weight of the stone carried by the goose on its feet.

3-4-1- Exploitation Phase

The possibility of safeguarding the groups is a prerequisite we have for the exploitation phase. We will find the weight of the stone that the goose stores in its feet, which is estimated to be between 5 and 25 g Through this Equation 8, we find the weight of the stone randomly for any iteration. This variable indicates the number of iterations.

$$TX_{it} = randi([5,25]) \tag{8}$$

Then, in Equation 9, we should find the time U_pBP_{it} needed to reach the earth when the stone falls. It's randomly between 1 and the number of dimensions for each iteration in the loop.

$$U_pBP_{it} = rand(1, dm) \tag{9}$$

In Equation 10, we find the time U_pBT_{it} when the object hits the ground and a sound is made and transmitted to the individual goose in the herd.

$$U_pBT_{it} = rand(1, dm) \tag{10}$$

In the next equation, discover the total time required for the sound to propagate and reach the individual goose in the flock throughout the iterations. As shown in Equation 11, the dimensions divide the total amount of time. To obtain the average time required, we divide the total time by 2. Equation 12 explains the steps.

$$UU = \frac{1}{dm}\sum_{j=1}^{dm} U_pB_{jt} \tag{11}$$

$$U_{UB} = \frac{UU}{2} \tag{12}$$

There is a random variable *rd* responsible for the distribution of the exploitation and exploration phases. The value of variable *pro* is randomly selected from the range [0, 1]. Consider the value of variable *pro* is greater than 0.2 and TX_{it} greater than or equal to 12. In Equation 13, U_pBP_{it} is multiplied by the square root of the TX_{it} divided by the object's acceleration at 9.81 m per square second, M/S^2. To protect and awaken the individual in his group, these equations should be worked out.

$$GGT = U_pBP_{it} * \frac{\sqrt{TX_{it}}}{9.81} \tag{13}$$

In Equation 14, to find the distance of sound travel ETU_{it}, it must be the speed of sound c_s in the air multiplied by the time of sound travel U_pBT_{it}. The speed of sound is 343.2 m/s in the air.

$$ETU_{it} = c_s U_p BT_{it} \quad (14)$$

In this step, we find EH_{it} the distance between the guard goose and another goose that is resting or feeding. In Equation 15, we use the distance of sound travel EHU_{it} multiplied by $\frac{1}{2}$ because we only need the time for the sound to travel and not the time for the sound to return.

$$EH_{it} = \frac{EHU_{it}}{2} \quad (15)$$

To resolve a new Y in the population. In other words, to wake up the individual in the flocks, we must find a $BestY_{it}$, as demonstrated in Equation 16. This equation is composed of the speed of the falling object GGT added to the distance of the Goose EH_{it} multiplied by the average of time squared UB.

$$Y_{it+1} = GGT + EH_{it} * UB^2 \quad (16)$$

On the contrary, if both variables are the weight of the stone TX_{it} and *pro*, one after the other less than 12 and less than or equal to 0.2. In Equation 17 to obtain the speed of a falling object GGT, multiply the time $U_p BP_{it}$ taken to arrive at the object by the weight of the stone TX_{it} divided by gravity. In addition, to determine the distance of sound travel ETU_{it} and the distance of the goose EH_{it} we dramatically used the previous Equations 14 and 15.

$$GGT = U_p BP_{it} * \frac{TX_{it}}{9.81} \quad (17)$$

In the other way, we find a new Y in the new mathematical equation. In Equation 18 all variables such as the speed of the falling object, distance of the goose, average of time, and coe are multiplied by each other in succession.

$$Y_{it+1} = GGT * EH_{it} * UB^2 * coe \quad (18)$$

In the exploitation phase, we used two equations to discover a new Y, for instance, Equations 16 and 18. These values of variables *pro* and TX_{it} determined which equation was performed.

3-4-2- Exploration Phase

In this stage, the goose awakens randomly, guided by the best position identified thus far, in order to regulate random activation or safeguard the individual. If a goose is not carrying stones on its feet, other members of the flock may awaken at random. Once one goose awakens, it begins calling out to alert and protect the rest of the flock. As outlined in earlier sections, when the value of the variable *rd* is less than 0.5, equations such as Equations 11 and 12 are applied. Additionally, a condition is checked to ensure that if the minimum time N_U exceeds the total time U_U, then the minimum time is reassigned to equal the total time. The parameter α varies from 2 to 0, decreasing sharply with each iteration of the loop. Finally, Equation 19 is employed to refine the value of a new Y within the search space.

$$\alpha = 2 - \left(\frac{2loop}{Max\ It}\right) \quad (19)$$

where, $Max\ It$ denotes the maximum number of iterations allowed. To guide the search process toward the solution most likely to be optimal, it is essential to compute the two parameters NU and alpha. The goose explores other individuals in the search space stochastically through *rdn(1, dm)*. Both NU and alpha play a key role in enhancing the search capability of GOOSE. In Equation 20, the minimum of time and alpha is multiplied by a random number and then added to the best position in the search space.

$$Y_{it+1} = randn(1, dm) * (NU * \alpha) + Best_{pos} \quad (20)$$

where, *dm* is the dimensionality of the problem, and $Best_{pos}$ denotes the *BestY*, in example, the best position identified in the search space. A detailed description of the GOOSE process can be found in Hamad et al. [43]. Figure 6 illustrates the flowchart of the Goose Algorithm, showing its main steps and decision processes.

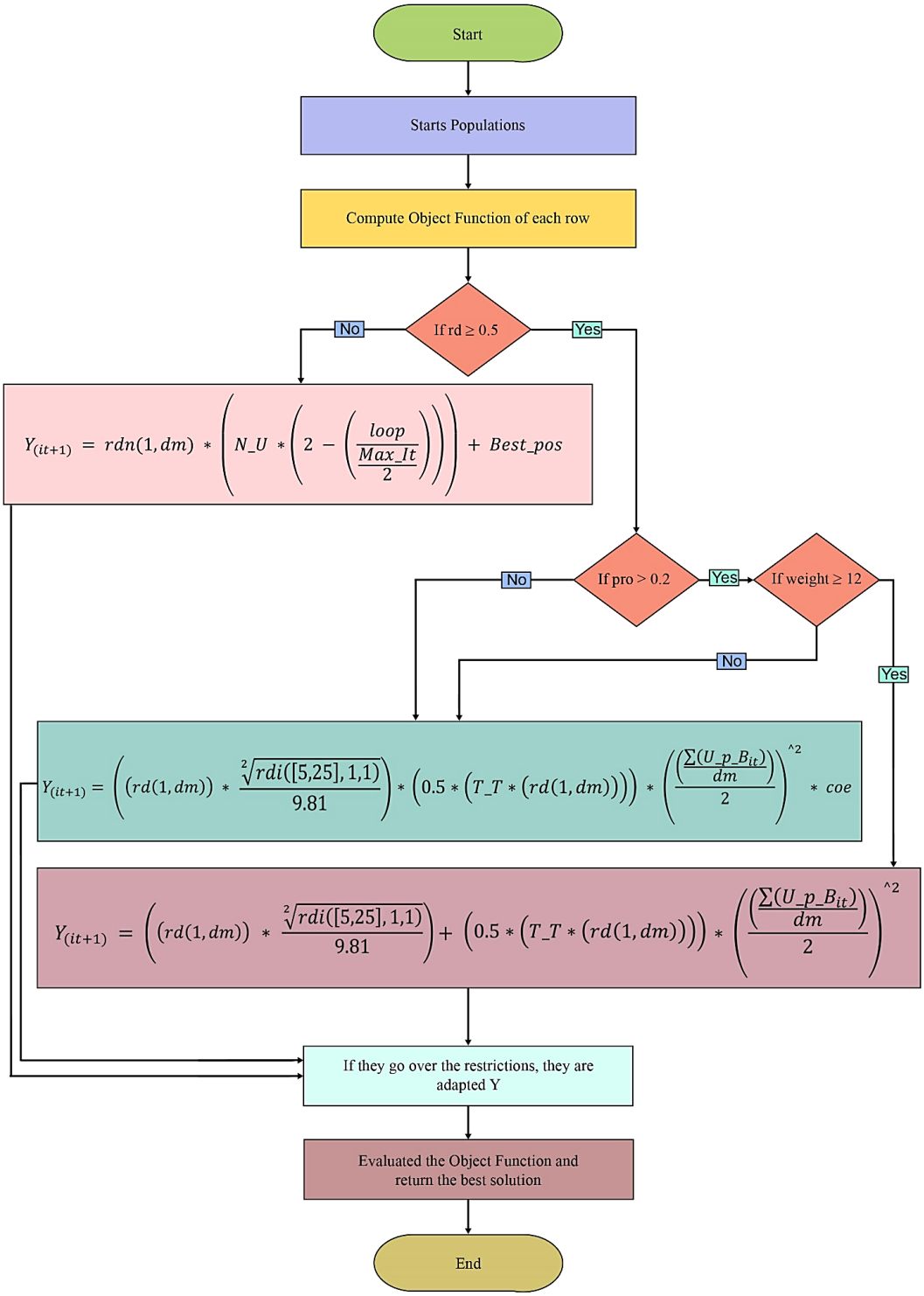

Figure 6. Flowchart of Goose Algorithm

3-5- Komodo Mlipir Algorithm

The Komodo Mlipir Algorithm (KMA), proposed by Suyanto et al. [44], is a modern metaheuristic inspired by the hunting and reproductive behaviors of Komodo dragons, as well as the Javanese term *mlipir* (which refers to walking safely along the roadside). In this algorithm, the population is divided into three categories of Komodo individuals: big males, a female, and small males. The process starts with the movement of big males, representing high-exploitation and low-exploration (HILE). Next, the female seeks solutions either by mating with the fittest big male or through parthenogenesis (exploration). Finally, the small males enhance diversity in the solution space using the *mlipir* movement. The Komodo population is split into three groups using Equations 21 and 22. The portion of the splitting Komodo groups is denoted as g using the interval (0-1). This portion is then used to split c number of Komodo into z big males, one female, and y small males. After that, the movement of each big male (kd_i) produces a new position (kp_i') expressed by Equations 23 and 24.

$$z = [(g-1)n] \tag{21}$$

$$y = c - z \tag{22}$$

$$kp_i' = kd_i + \sum_{j=1}^{z} w_{ij}, \quad where\ j \neq i \tag{23}$$

$$w_{ij} = \begin{cases} r_1(kd_j - kd_i), & if \quad f(kd_j) < f(kd_i) \quad or \quad r_2 < 0.5 \\ r_1(kd_i - kd_j), & otherwise \end{cases} \tag{24}$$

The movement of Komodo dragons, as defined in Equation 23, ensures that the q highest-quality big males survive into the next generation. The female either mates with the top-performing big male using Equations 25 and 26 or reproduces through parthenogenesis using Equation 27.

$$kp_{il}' = r_l.kd_{il} + (1 - r_l).kd_{jl} \tag{25}$$

$$kp_{jl}' = r_l.kd_{jl} + (1 - r_l).kd_{il} \tag{26}$$

$$kp_{ij}' = kd_{ij} + (2r - 1) \propto |ub_j - lb_j| \tag{27}$$

$$\{kd_{i1}, kd_{i2}, \ldots, kd_{im}\} \rightarrow \{kp_{i1}', kp_{i2}', \ldots, kp_{im}'\} \tag{28}$$

Here, kd_{il} and kd_{jl} represent the *l*-th dimension of the *i*-th and the *j*-th winning big males and the female Komodos, respectively. The mating process generates two offspring, which are the kp_{il}' and kp_{jl}'. In parthenogenesis, a small value is added to each dimension of the female, as shown in Equation 28. The movement of small males is described by Equations 29 and 30, which helps maintain their positions for survival in the next generation. Finally, the population size *c* is updated using Equation 31. A detailed description of the KMA process can be found in Suyanto et al. [44].

$$w_{ij} = \begin{cases} \sum_{l=1}^{m} r_1(kd_{jl} - kd_{il}), & if \quad r_2 < d \\ 0, & otherwise \end{cases} \tag{29}$$

$$kp_i' = kd_i + \sum_{j=1}^{z} w_{ij}, where\ j \neq i \tag{30}$$

$$c' = \begin{cases} c - a, & if\ \delta\ f_1 > 0\ \ and\ \ \vartheta\ f_2 > 0 \\ c + a, & if\ \delta\ f_1 = 0\ \ and\ \ \vartheta\ f_2 = 0 \end{cases} \tag{31}$$

3-6- Evaluation Metrics

To assess the performance of the proposed multimodal emotion recognition framework, we employed a set of standard evaluation metrics widely used in classification tasks [45]. These metrics provide complementary perspectives on model effectiveness by measuring not only the overall accuracy but also the balance between correct predictions across different emotion categories. Since emotion recognition often involves class imbalance and overlapping affective states, relying solely on accuracy can be misleading. Therefore, metrics such as precision, recall, and F1-score are included to ensure more.

1) Accuracy

Accuracy refers to the proportion of correctly classified samples compared to the total dataset. It indicates how many predictions were made correctly relative to the overall number of data points. This measure is calculated using Equation 32.

$$Accuracy(h) = \frac{1}{|X|}\sum_{x \in X}[h(x) = y(x)] \tag{32}$$

2) Precision

Precision refers to the proportion of harmful programs correctly identified out of all applications classified as harmful. This metric takes into account both true positives and false positives. A higher precision value is achieved when the number of false positives is minimal. The computation is given in Equation 33.

$$Precision\ (h) = \frac{\sum_{j=1}^{l} t_{Pj}}{\sum_{j=1}^{l}(t_{Pj} + f_{Pj})} \tag{33}$$

Here, t_p corresponds to the count of true-positive determinations, and f_p represents the count of false-positive identifications.

3) Recall

Recall, also known as True Positive Rate (TPR), represents the ratio of precisely predicted values to the total number of records for each class. It assesses how well the model identifies positive instances from the entire dataset and the Equation 34 is utilized to compute recall.

$$Recall\ (h) = \frac{\sum_{j=1}^{l} t_{Pj}}{\sum_{j=1}^{l}(t_{Pj} + f_{nj})} \tag{34}$$

t_p denotes the number of true positives identified by the model, whereas f_n represents the number of false negatives produced by the algorithm.

4) F1-score

The F1 score represents the harmonic average of precision and recall, and its computation is expressed in Equation 35.

$$F1 - score = \frac{2 \times True\ Positives}{2 \times True\ Positives + False\ Positives + False\ Negatives} \tag{35}$$

5) Confusion Matrix

The confusion matrix is a valuable tool for evaluating the performance of a classification model in both binary and multi-class scenarios. It provides important insights into metrics such as accuracy, precision, and recall.

4- Results and Discussion

This section presents the experimental findings of the proposed hybrid framework for multimodal emotion recognition, followed by an in-depth discussion of the results. The evaluation was carried out across three settings—audio modality, text modality, and the fusion of both modalities—in order to assess the contribution of each component and the effectiveness of the overall system. Performance metrics including precision, recall, F1-score, and accuracy were used to evaluate classification quality across six emotion categories: *neutral, happy, sad, angry, excited,* and *frustrated*. The results are first reported for unimodal experiments (audio and text separately) to establish baseline performance, after which the fusion results using the Komodo Mlipir Algorithm (KMA) are presented. Finally, a comparative analysis across modalities highlights key strengths, weaknesses, and misclassification patterns, providing insight into the advantages of the proposed multimodal integration.

Table 3. Audio

Emotion	Audio low				Audio high			
	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
Neutral	1.0000	1.0000	1.00	2	0.0000	0.0000	0.0000	2
Happy	0.7143	1.0000	0.83	5	0.4444	0.8000	0.5714	5
Sad	1.0000	0.6000	0.75	5	0.6000	0.6000	0.6000	5
Angry	1.0000	1.0000	1.00	4	0.5000	0.2500	0.3333	4
Excited	0.8571	1.0000	0.92	6	0.5714	0.6667	0.6154	6
Frustrated	0.8750	0.7778	0.82	9	0.3750	0.3333	0.3529	9
Accuracy			0.87	31			0.4839	31
Macro avg	0.9077	0.8963	0.89	31	0.4151	0.4417	0.4122	31
Weighted avg	0.8900	0.8710	0.87	31	0.4524	0.4839	0.4535	31

The result for the audio is shown in Table 3. For audio low, the precision obtained by Neutral, Happy, Sad, Angry, Excited, Frustrated are 100%, 71.43%, 100%, 100%, 85.71%, 87.50%. For audio high, the precision obtained by Neutral, Happy, Sad, Angry, Excited, Frustrated are 0%, 44.44%, 60%, 50%, 57.14%, 37.50%. For audio low, the recall obtained by Neutral, Happy, Sad, Angry, Excited, Frustrated are 100%, 100%, 60%, 100%, 100%, 77.78%. For audio high, the recall obtained by Neutral, Happy, Sad, Angry, Excited, Frustrated are 0%, 80%, 60%, 25%, 66.67%, 33.33%. For audio low, the F1-Score obtained by Neutral, Happy, Sad, Angry, Excited, Frustrated are 100%, 83%, 75%, 100%, 92%, 82%. For audio high, the recall obtained by Neutral, Happy, Sad, Angry, Excited, Frustrated are 0%, 57.14%, 60%, 33.33%, 61.54%, 35.29%. The results for the audio modality indicate a strong performance at the low feature level, achieving an

overall accuracy of 87%. Emotions such as *neutral*, *angry*, and *excited* were recognized with perfect or near-perfect scores (precision and recall above 0.85). For example, *angry* achieved precision and recall of 1.00, while *excited* reached an F1-score of 0.92. However, recognition performance dropped considerably at the high feature level, with overall accuracy decreasing to 48.4%. In this setting, emotions such as *frustrated* and *angry* were poorly detected, with F1-scores of 0.35 and 0.33, respectively. This suggests that higher-level audio features may introduce noise or overlap, making it harder for the classifier to discriminate subtle affective differences.

Table 4. Text

Emotion	Precision	Recall	F1-Score	Support
Neutral	0.0000	0.0000	0.0000	2
Happy	0.7143	1.0000	0.8333	5
Sad	0.6667	0.4000	0.5000	5
Angry	0.5000	0.5000	0.5000	4
Excited	0.6667	0.6667	0.6667	6
Frustrated	0.5455	0.6667	0.6000	9
Accuracy			0.6129	31
Macro avg	0.5155	0.5389	0.5167	31
Weighted avg	0.5746	0.6129	0.5828	31

The result for the text is shown in Table 4. The precision obtained by Neutral, Happy, Sad, Angry, Excited, Frustrated are 0%, 71.43%, 66.67%, 50%, 66.67%, 54.55%. The recall obtained by Neutral, Happy, Sad, Angry, Excited, Frustrated are 0%, 100%, 40%, 50%, 66.67%, and 66.67%. The F1-Score obtained by Neutral, Happy, Sad, Angry, Excited, Frustrated are 0%, 83.33%, 50%, 50%, 66.67% and 60%. For the text-only modality, the system achieved an accuracy of 61.3%. Certain emotions like *happy* were recognized reliably, with an F1-score of 0.83, while *excited* and *frustrated* showed moderate performance (0.67 and 0.60 F1, respectively). On the other hand, *neutral* was not detected at all, yielding zero scores across precision, recall, and F1. Similarly, *sad* and *angry* obtained relatively low F1-scores (0.50 each). These results highlight a key limitation of text-based recognition: while lexical cues may capture some affective signals, they are insufficient to consistently identify emotions that rely on prosody, intensity, or contextual nuance.

Table 5. Fusion (final)

Emotion	Precision	Recall	F1-Score	Support
Neutral	1.0000	1.0000	1.0000	1
Happy	0.7500	1.0000	0.8571	3
Sad	1.0000	0.5000	0.6667	2
Angry	1.0000	1.0000	1.0000	2
Excited	1.0000	1.0000	1.0000	3
Frustrated	0.8000	0.8000	0.8000	5
Accuracy			0.8750	16
Macro avg	0.9250	0.8833	0.8873	16
Weighted avg	0.8906	0.8750	0.8690	16

The result for the modalities using KMA for late fusion is shown in Table 5. The precision obtained by Neutral, Happy, Sad, Angry, Excited, Frustrated are 100%, 75%, 100%, 100%, 100%, 80%. The recall obtained by Neutral, Happy, Sad, Angry, Excited, Frustrated are 100%, 100%, 50%, 100%, 100%, and 80%. The F1-Score obtained by Neutral, Happy, Sad, Angry, Excited, Frustrated are 100%, 85.71%, 66.67%, 100%, 100% and 80%. The integration of modalities using KMA for late fusion significantly improved performance, with overall accuracy reaching 87.5%. Several emotions were recognized with perfect precision and recall, including *neutral*, *angry*, and *excited*. *Happy* also showed strong performance with an F1-score of 0.86. Although *sad* was partially misclassified (recall = 0.50), its precision remained perfect (1.00), suggesting that when the system predicted sadness, it was highly reliable. *Frustrated* maintained a balanced performance with an F1-score of 0.80. These results confirm the effectiveness of the hybrid approach in combining complementary cues from audio and text, while mitigating the weaknesses observed in unimodal settings.

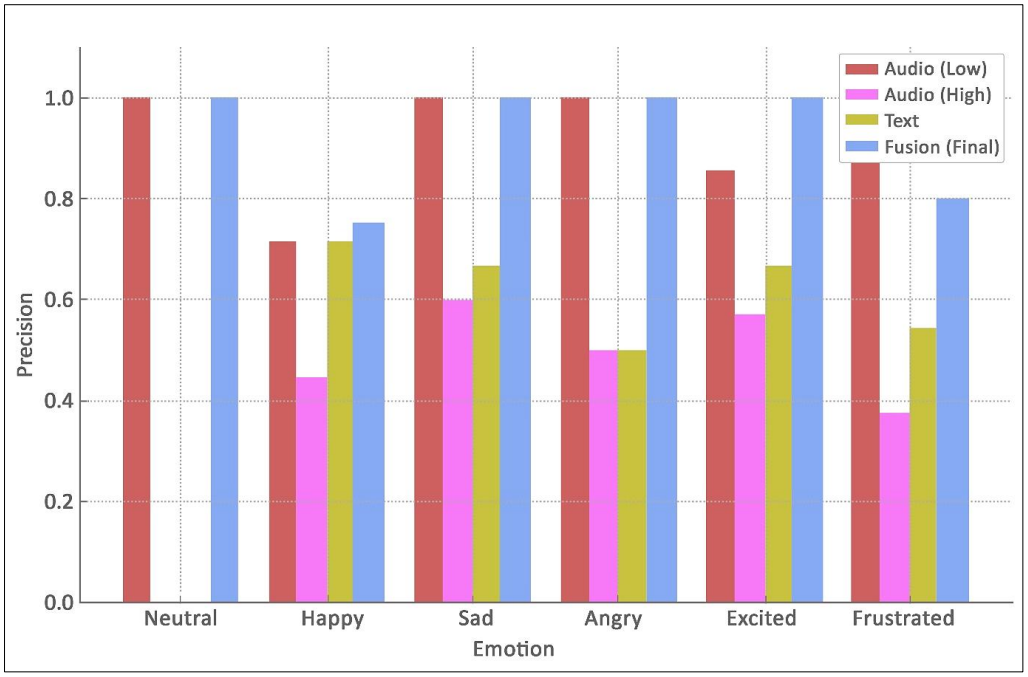

Figure 7. Per-emotion precision performance comparison across modalities

Figure 7 shows per-emotion precision comparison across four modalities: Audio (Low), Audio (High), Text, and Final Fusion. The results show that fusion consistently outperforms unimodal approaches, achieving perfect recognition for Neutral, Sad, Angry, and Excited, while also improving performance for Happy and Frustrated. Audio (Low) performs strongly for most emotions but is slightly weaker on Happy, whereas Text is effective for Happy and Sad but fails on Neutral. Audio (High) yields the weakest results overall. These findings confirm the advantage of multimodal fusion in providing robust and balanced emotion recognition.

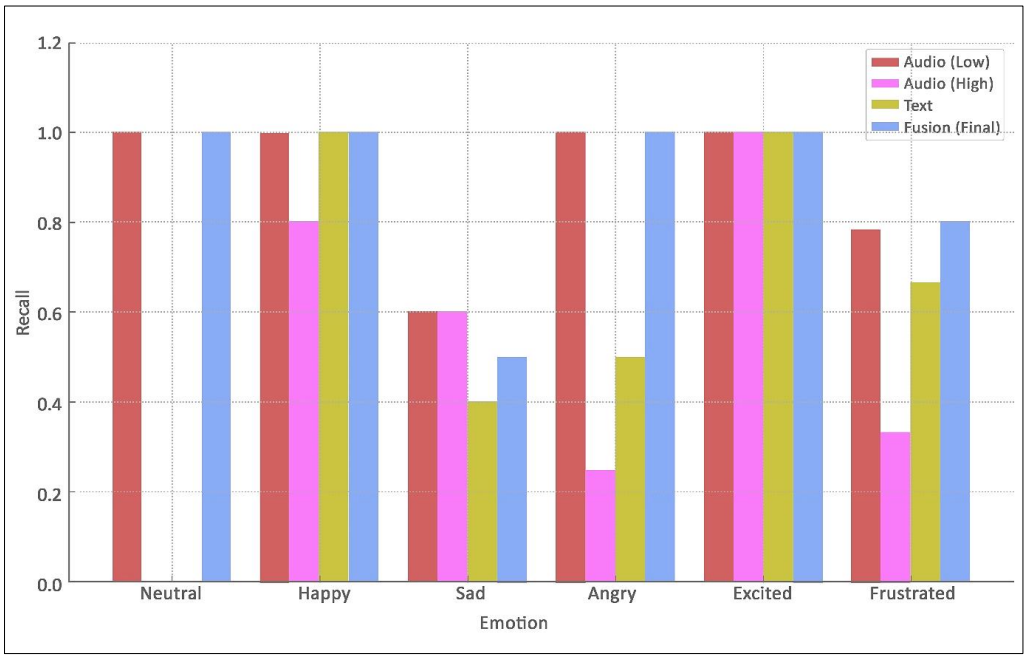

Figure 8. Per-emotion recall performance comparison across modalities

Figure 8 shows per-emotion recall comparison across four modalities: Audio (Low), Audio (High), Text, and Final Fusion. Fusion consistently achieves high recall for most emotions, with perfect recognition of Neutral, Happy, Angry, and Excited. Audio (Low) also performs strongly, especially for emotions expressed with clear prosodic cues such as Angry and Neutral. Text shows high recall for Happy and Excited but fails for Neutral and performs weakly for Sad. Audio (High) generally underperforms, suggesting that higher-level acoustic features are less reliable. These results

highlight the complementary role of multimodal fusion while also revealing that fusion may not always outperform unimodal methods for certain subtle emotions like Sad.

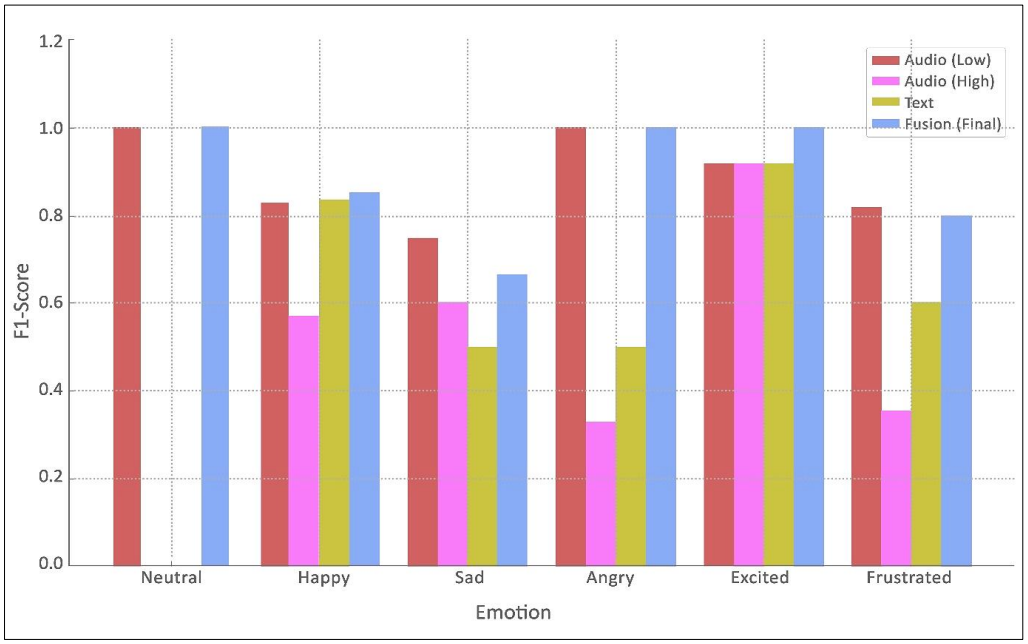

Figure 9. Per-emotion F1-Score performance comparison across modalities

Figure 9 shows per-emotion F1-score comparison across four modalities: Audio (Low), Audio (High), Text, and Final Fusion. Fusion achieves the best or tied-best performance for Neutral, Happy, Angry, and Excited, confirming the benefit of combining modalities. Audio (Low) remains a strong standalone modality, particularly for Neutral, Angry, and Frustrated, while Text contributes effectively to Happy and Excited. Audio (High) consistently underperforms. For Sad and Frustrated, fusion does not surpass the best unimodal result, suggesting that certain emotions are more modality-dependent and less improved by fusion.

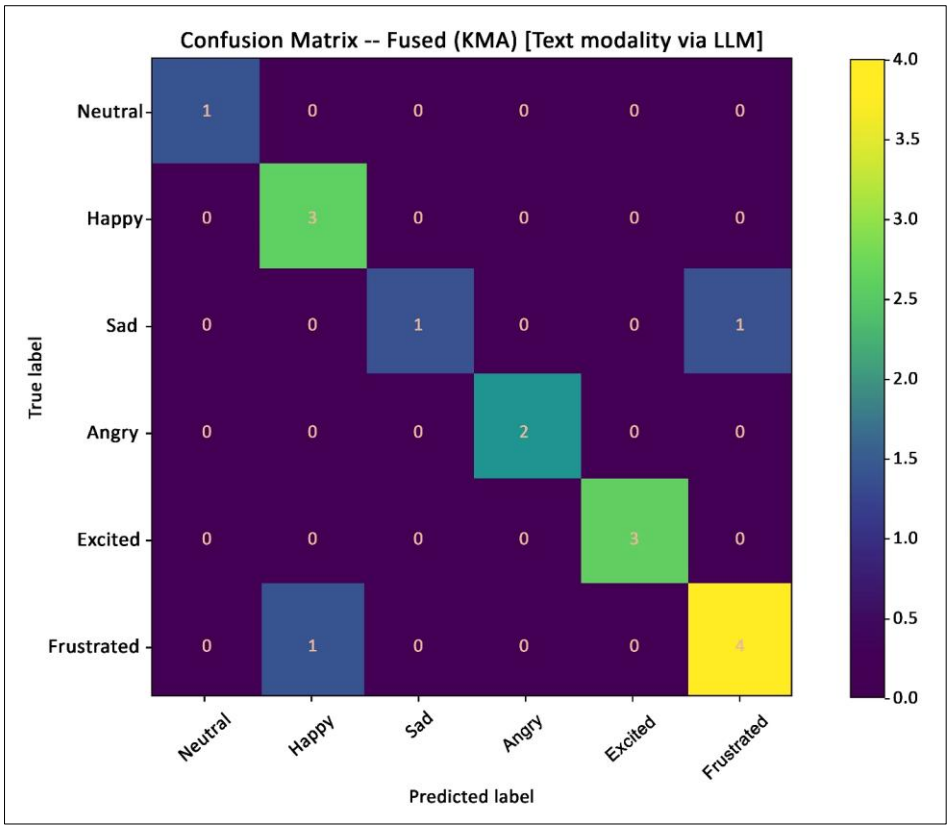

Figure 10. Confusion matrix

The performance of the proposed multimodal emotion recognition framework was evaluated using a confusion matrix, as shown in Figure 10. The results demonstrate the effectiveness of the fused approach—combining LLM-based textual representations with the Komodo Mlipir Algorithm (KMA) for late fusion—across six emotion categories: neutral, happy, sad, angry, excited, and frustrated. The framework successfully recognized most emotions with high accuracy. Specifically: Neutral was correctly classified with no misclassification (1/1). Happy achieved strong recognition (3/4), with one instance misclassified as *frustrated*. Sad showed mixed results (1/2), with one sample confused with *frustrated*. Angry was classified consistently (2/2), indicating that the model effectively captured distinctive textual cues for anger. Excited obtained full recognition (3/3), showing that the model was sensitive to lexical and contextual markers of high-arousal positive states. Frustrated achieved the best recognition performance (4/5), though one instance was misclassified as *happy*.

The main misclassification patterns occurred between sad and frustrated, as well as between happy and frustrated. This aligns with psychological findings that frustration shares linguistic and contextual similarities with both sadness (low valence, negative tone) and happiness (expressions of intensity or exclamatory forms that overlap with excited/happy tones). These overlaps may explain the observed confusion in the model. The results highlight several key insights: 1) Strength of LLMs in text-based emotion recognition: Emotions such as *angry* and *excited*, which often carry strong lexical markers (e.g., exclamation words, intensifiers), were recognized with high precision. This supports the effectiveness of LLMs in capturing contextual semantics. 2) Challenges in distinguishing subtle negative emotions: The overlap between *sad* and *frustrated* suggests that fine-grained distinctions in negative affect remain challenging for text-only models, particularly when contextual cues are limited. 3) Effectiveness of fusion via KMA: Despite some misclassifications, the overall distribution shows that the KMA-based late fusion contributed to stable recognition across multiple categories, reducing noise from ambiguous cases. These results indicate that while the proposed hybrid framework performs strongly in recognizing emotions with distinct textual signals, future improvements are required for differentiating subtle or overlapping affective states. Incorporating additional modalities such as acoustic features (tone, pitch, prosody) or visual signals (facial expressions) could enhance differentiation between emotions such as sadness and frustration.

In addition to classification performance, the interpretability of the proposed hybrid framework is a key methodological strength, enabling a multi-layered analysis of the model's decisions at the modality, feature, and token levels. This hybrid approach combines both model-based and post-hoc interpretability methods to provide a comprehensive understanding of the model's decision-making process [32, 46].

At the highest level—the "multimodal decision" itself—the Komodo Mlipir Algorithm (KMA) provides a direct and quantitative measure of modality importance. The KMA-optimized weights, which determine the contribution of the audio and text streams during late fusion, function as a learned, global explanation for the model's fusion policy. This automated weighting mechanism is analogous to perturbation-based interpretability methods, such as occlusion sensitivity, as the resulting weights (e.g., w_{audio} vs. w_{text}) explicitly quantify the model's learned reliance on each modality for accurate emotion classification.

This is further complemented at the feature-level by the Binary Artificial Hummingbird Algorithm (BAHA). As a wrapper-based method, BAHA's role is dual-purpose: beyond optimizing performance, it provides intrinsic interpretability. The binarization of the core Artificial Hummingbird Algorithm is designed to solve discrete selection problems [32]. The final binary vector produced by BAHA (Eq. 1) serves as a direct feature importance map, revealing the minimal, most discriminative subset of acoustic features (e.g., specific MFCCs, pitch, energy) that the model found most relevant, thereby simplifying the model and enhancing transparency.

Furthermore, the framework facilitates token-level interpretability for the textual modality, which is processed by the Sentence Transformer [47] (Section 3.2.). While direct self-attention visualization from the Transformer architecture offers a preliminary view of token-to-token interactions, a more robust and faithful explanation of the model's decision-making can be extracted using post-hoc attribution methods [48]. The LLM component is amenable to gradient-based Explainable AI (XAI) techniques, such as Layer-wise Relevance Propagation (LRP) [49] or Integrated Gradients [50]. These methods, which have proven effective in extracting high-resolution patterns from complex signal data [51], can generate saliency maps that attribute the classification score back to the individual input tokens. This allows for a granular analysis of which specific words or phrases (e.g., "so stupid" in Figure 2) contribute most significantly to a given emotion prediction, such as 'frustrated', thereby completing the framework's hierarchical interpretability from the multimodal fusion down to the raw input.

Table 6 shows comparison between our methods and other emotion recognition methods on IEMOCAP dataset. The experimental results showed that the proposed method was more accurate than state-of-the-art ones in terms of detecting emotions.

Table 6. Comparison between proposed method and other emotion recognition methods on IEMOCAP dataset

Author	Methods	Modality	Accuracy
Hong et al. [16]	AER LLM	Text	56%
Michael et al. [12]	Genetic Algorithm	Audio, Text	66.67%
Li et al. [52]	Shared Autoencoder, Bi-LSTM	Audio, Text	70%
Zhang et al. [25]	DialogueLLM	Text	70.48%
Zou et al. [53]	Bi-LSTM, Cross Modal Transformer	Audio, Text, Video	72.27%
Wang et al. [54]	RNN	Audio	72.30%
Chamishka et al. [55]	BiDialogRNN with GRUs	Audio	73.81%
Chauhan et al. [56]	CNNEncoder, Multiheadattention Network	Audio	73.81%
Singh et al. [57]	Hierarchical DNN	Audio, Text	74.50%
Zhang et al. [58]	Encoder-Decoder Network	Audio, Text, Video	81.20%
Yoon [59]	CrossModal Translation	Audio, Text, Video	83.20%
Braunschweiler et al. [60]	CNN-Bi_LSTM with attention	Audio, Text	85.50%
Chaudhari et al. [11]	GCM-Net	Audio, Text, Video	85.66%
Proposed Method	LLM, BAHA, GOOSE, KMA	Audio, Text	87.50%

We have integrated the model into the Gradio graphical user interface once it is prepared. This integration is crucial because it provides an accessible multimodal emotion data and understandable visual representation of the model's training output. Users can input and see the model predictions instantly in real time with the Gradio interface, which facilitates intuitive interaction. This visual aid makes it easy to test and understand how well the model predicts, making it essential for ongoing validation and real-world use [61].

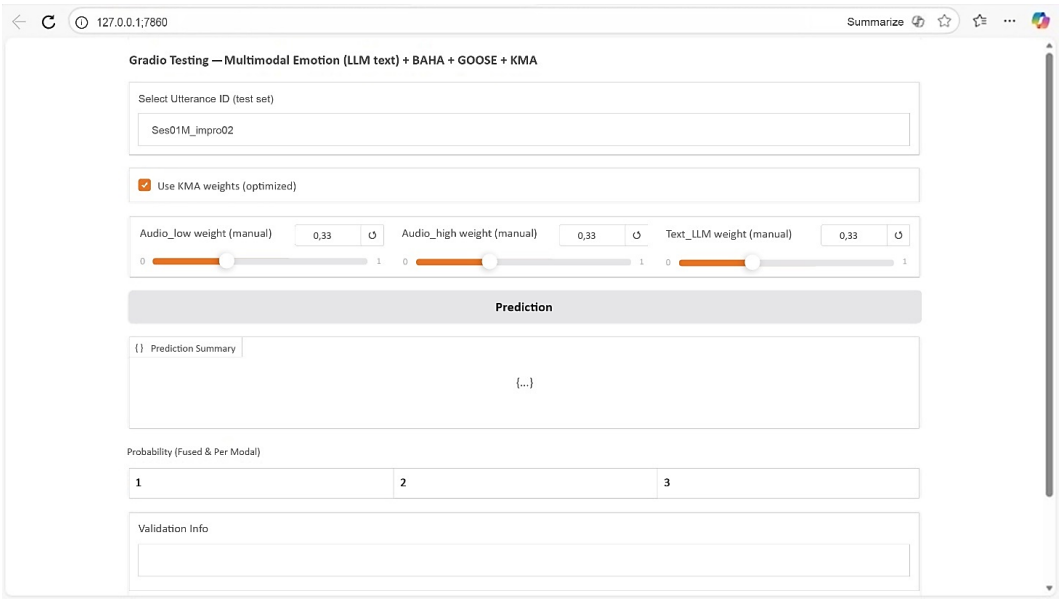

Figure 11. Gradio interface for multimodal emotion recognition

Figure 11 shows a Gradio interface for multimodal emotion recognition. The displayed Gradio interface serves as an interactive testing platform for a multimodal emotion recognition system. It allows users to evaluate how well the model predicts emotions from a combination of three modalities — audio low-level features, audio high-level features, and text embeddings generated using a LLM. The interface is designed to make the model's behavior transparent and easy to explore, showing how each modality contributes to the final emotion prediction. At the top, users can select an utterance ID from the test dataset. This ID represents a single spoken segment, including both its audio and textual transcript. Once an ID is chosen, the system retrieves the relevant data and prepares it for prediction. The user can then choose to apply KMA-optimized weights, which were determined automatically during model training to provide the best balance among modalities. Alternatively, users can manually adjust the contribution of each modality using three sliders for audio_low, audio_high, and text_llm, allowing them to experiment with different weighting schemes. When the Prediction button is clicked, the system processes the selected utterance through each modality's trained classifier. It then fuses the results based on the chosen weighting method; either the optimized KMA weights or user-defined manual weights. The interface displays the final predicted emotion (fused), along with individual predictions from each

modality, enabling users to compare and understand how the fusion improves accuracy. Below the summary section, the interface also presents a probability table that lists each possible emotion along with its predicted confidence scores from the fused output and from each modality separately. This detailed view helps users see which emotion classes are most strongly supported by each data source. At the bottom, the Validation Info section shows additional details, such as the Macro-F1 score from validation using KMA weights, offering insight into the model's overall performance. This Gradio UI provides a clear, hands-on way to visualize and interpret the results of a complex multimodal fusion model. It combines transparency, flexibility, and interactivity — letting users test utterances, adjust fusion weights, and observe how different modalities jointly determine the final emotion classification. Figure 12 shows implementation of gradio interface for multimodal emotion recognition.

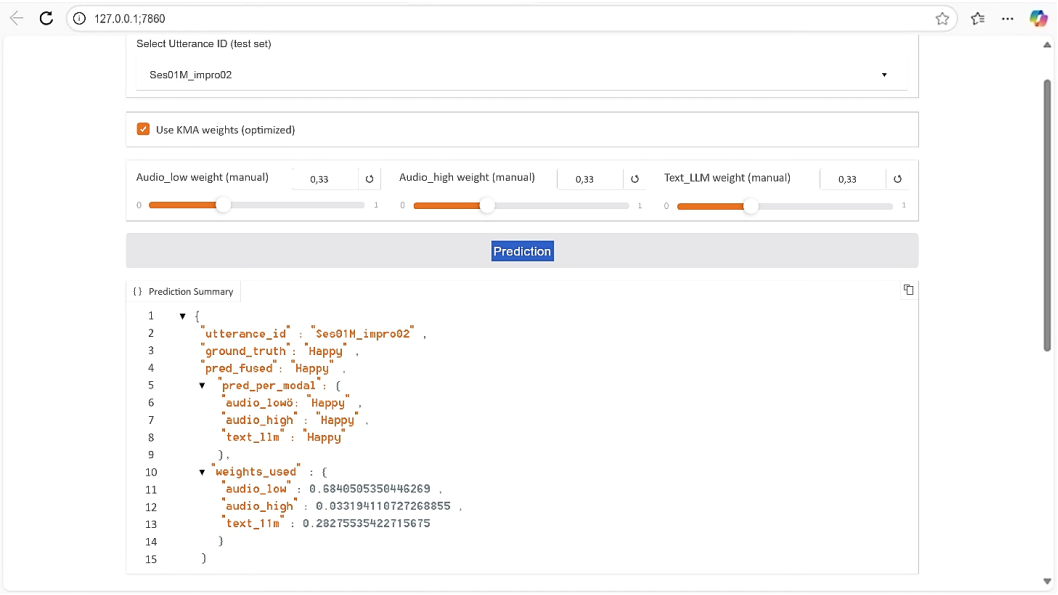

Figure 12. Implementation of Gradio interface for multimodal emotion recognition

In summary, compared to existing real-time emotion recognition systems, the proposed LLM–metaheuristic hybrid offers a favorable accuracy–latency trade-off. It outperforms many real-time systems in accuracy due to its sophisticated multimodal integration, yet remains within a feasible latency range for interactive use. With recommended optimizations (distillation, parallel processing, etc.), it can match the inference speed of simpler models. This positions our framework as a promising solution for scenarios that demand both high recognition performance and real-time responsiveness – a combination that is increasingly sought after in areas like intelligent user interfaces, mental health monitoring, and adaptive learning environments. The results and analysis indicate that, through careful system-level engineering, we can bridge the gap between advanced deep learning models and the stringent timing requirements of real-world emotion-aware applications.

5- Conclusion

In conclusion, we presented a novel multimodal emotion recognition approach that synergistically integrates an LLM with metaheuristic optimization. The proposed hybrid framework effectively addresses key challenges in MER by combining LLM-based semantic feature extraction with BAHA-driven feature selection, GA-based hyperparameter tuning, and KMA-mediated late fusion. Our experimental results on the IEMOCAP dataset demonstrate that this integrated strategy yields significant performance gains: the model achieved 87.5% accuracy across six emotion classes, surpassing existing state-of-the-art methods. These findings confirm that leveraging LLMs alongside tailored metaheuristic algorithms can capture subtle affective nuances and enhance the overall reliability of emotion classification systems.

Despite the promising results, this work has some limitations that point to avenues for future research. First, our evaluation was limited to a single benchmark (IEMOCAP); further studies should validate the framework on additional datasets and in real-world scenarios to ensure its generalizability. Second, the current implementation focuses on two modalities (audio and text) and does not incorporate visual signals such as facial expressions, which are important for a more complete understanding of emotion; integrating such modalities could further improve performance in distinguishing closely related emotions. Moreover, the introduction of LLM components increases model complexity; future efforts may explore more efficient or distilled LLM architectures to reduce computational overhead. Finally, research into enhancing the interpretability of the hybrid model (for example, by analyzing feature importance or decision rationale) would be valuable for deploying emotionally intelligent systems in sensitive applications. By addressing these aspects, the proposed framework can be extended and refined to build more robust, versatile, and explainable multimodal emotion recognition systems in the future.

6- Declarations

6-1- Author Contributions

Conceptualization, A.M., M.T.U., and C.A.R.; methodology, A.M.; software, A.M., A.S., and L.A.R.; validation, A.M., M.T.U., A.S., L.R.A., R.W., A.F., and C.A.R.; formal analysis, A.M., M.T.U., A.S., and C.A.R.; investigation, A.M., M.T.U., and C.A.R.; resources, A.M., M.T.U., A.S., and L.A.R.; data curation, A.M., M.T.U., A.F., and C.A.R.; writing—original draft preparation, A.M., M.T.U., L.A.R., and R.W.; writing—review and editing, A.M., M.T.U., A.S., L.R.A., R.W., A.F., and C.A.R.; visualization, A.M., M.T.U., and A.S.; supervision, M.T.U, A.F., and C.A.R.; project administration, M.T.U. and C.A.R.; funding acquisition, M.T.U. and C.A.R. All authors have read and agreed to the published version of the manuscript.

6-2- Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6-3- Funding and Acknowledgments

This research project is supported by Second Century Fund (C2F), Chulalongkorn University, Thailand. We gratefully appreciate this support.

6-4- Institutional Review Board Statement

Not applicable.

6-5- Informed Consent Statement

Not applicable.

6-6- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

7- References

[1] Younis, E. M. G., Mohsen, S., Houssein, E. H., & Ibrahim, O. A. S. (2024). Machine learning for human emotion recognition: a comprehensive review. Neural Computing and Applications, 36(16), 8901–8947. doi:10.1007/s00521-024-09426-2.

[2] Guo, R., Guo, H., Wang, L., Chen, M., Yang, D., & Li, B. (2024). Development and application of emotion recognition technology — a systematic literature review. BMC Psychology, 12(1), 95. doi:10.1186/s40359-024-01581-4.

[3] Ramaswamy, M. P. A., & Palaniswamy, S. (2024). Multimodal emotion recognition: A comprehensive review, trends, and challenges. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 14(6), 1563. doi:10.1002/widm.1563.

[4] Khare, S. K., Blanes-Vidal, V., Nadimi, E. S., & Acharya, U. R. (2024). Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. Information Fusion, 102, 102019. doi:10.1016/j.inffus.2023.102019.

[5] Hazmoune, S., & Bougamouza, F. (2024). Using transformers for multimodal emotion recognition: Taxonomies and state of the art review. Engineering Applications of Artificial Intelligence, 133, 108339. doi:10.1016/j.engappai.2024.108339.

[6] Kalateh, S., Estrada-Jimenez, L. A., Nikghadam-Hojjati, S., & Barata, J. (2024). A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges. IEEE Access, 12, 103976–104019. doi:10.1109/ACCESS.2024.3430850.

[7] Zhang, S., Yang, Y., Chen, C., Zhang, X., Leng, Q., & Zhao, X. (2024). Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. Expert Systems with Applications, 237, 121692. doi:10.1016/j.eswa.2023.121692.

[8] Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E., & Azam, S. (2024). A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. IEEE Access, 12, 26839–26874. doi:10.1109/ACCESS.2024.3365742.

[9] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2025). A Comprehensive Overview of Large Language Models. ACM Transactions on Intelligent Systems and Technology, 16(5), 1–72. doi:10.1145/3744746.

[10] Pan, B., Hirota, K., Jia, Z., Zhao, L., Jin, X., & Dai, Y. (2023). Multimodal emotion recognition based on feature selection and extreme learning machine in video clips. Journal of Ambient Intelligence and Humanized Computing, 14(3), 1903–1917. doi:10.1007/s12652-021-03407-2.

[11] Chaudhari, P., Kumar, A., Raghaw, C. S., Rehman, M. Z. U., & Kumar, N. (2024). GCM-Net: Graph-enhanced cross-modal infusion with a metaheuristic-driven network for video sentiment and emotion analysis. arXiv Preprint, arXiv:2410.12828. doi:10.48550/arXiv.2410.12828.

[12] Michael, S., & Zahra, A. (2024). Multimodal speech emotion recognition optimization using genetic algorithm. Bulletin of Electrical Engineering and Informatics, 13(5), 3309–3316. doi:10.11591/eei.v13i5.7409.

[13] Mukta, M. S. H., Ahmad, J., Zaman, A., & Islam, S. (2024). Attention and Meta-Heuristic Based General Self-Efficacy Prediction Model from Multimodal Social Media Dataset. IEEE Access, 12, 36853–36873. doi:10.1109/ACCESS.2024.3373558.

[14] Daneshfar, F., Kabudian, S. J., & Neekabadi, A. (2020). Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and Gaussian elliptical basis function network classifier. Applied Acoustics, 166, 107360. doi:10.1016/j.apacoust.2020.107360.

[15] Dutta, S., & Ganapathy, S. (2025). LLM supervised Pre-training for Multimodal Emotion Recognition in Conversations. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 1–5. doi:10.1109/ICASSP49660.2025.10889998.

[16] Hong, X., Gong, Y., Sethu, V., & Dang, T. (2025). AER-LLM: Ambiguity-aware Emotion Recognition Leveraging Large Language Models. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 1–5. doi:10.1109/ICASSP49660.2025.10888198.

[17] Kim, E. H., Lim, M. J., & Shin, J. H. (2025). MMER-LMF: Multi-Modal Emotion Recognition in Lightweight Modality Fusion. Electronics (Switzerland), 14(11), 2139. doi:10.3390/electronics14112139.

[18] Li, Z., Lu, C., Xu, X., Zhang, K., Gu, Y., Li, B., Zong, Y., & Zheng, W. (2025). Enhancing Task-Specific Feature Learning with LLMs for Multimodal Emotion and Intent Joint Understanding. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 1–2. doi:10.1109/ICASSP49660.2025.10890555.

[19] Lu, H., Chen, J., Liang, F., Tan, M., Zeng, R., & Hu, X. (2025). Understanding Emotional Body Expressions via Large Language Models. Proceedings of the AAAI Conference on Artificial Intelligence, 39(2), 1447–1455. doi:10.1609/aaai.v39i2.32135.

[20] Teng, S., Liu, J., Sun, H., Chai, S., Tateyama, T., Lin, L., & Chen, Y. W. (2025). Enhanced Multimodal Depression Detection with Emotion Prompts. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 1–5. doi:10.1109/ICASSP49660.2025.10889035.

[21] Xu, X., Lu, C., Li, Z., Liu, Y., Ma, Y., Luo, J., Zong, Y., & Zheng, W. (2025). Reliable Learning from LLM Features for Multimodal Emotion and Intent Joint Understanding. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 1–2. doi:10.1109/ICASSP49660.2025.10888958.

[22] Yacoubi, I., Ferjaoui, R., Djeddi, W. E., & Khalifa, A. Ben. (2025). Advancing Emotion Recognition through LLaMA3 and LoRA Fine-Tuning. 22nd IEEE International Multi-Conference on Systems, Signals and Devices, SSD 2025, 348–353. doi:10.1109/SSD64182.2025.10989922.

[23] Yang, Y., Dong, X., & Qiang, Y. (2025). MSE-Adapter: A Lightweight Plugin Endowing LLMs with the Capability to Perform Multimodal Sentiment Analysis and Emotion Recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 39(24), 25642–25650. doi:10.1609/aaai.v39i24.34755.

[24] Zhang, S., Hu, Y., Yi, X., Nanayakkara, S., & Chen, X. (2025). IntervEEG-LLM: Exploring EEG-Based Multimodal Data for Customized Mental Health Interventions. WWW Companion 2025 - Companion Proceedings of the ACM Web Conference 2025, 2320–2326. doi:10.1145/3701716.3717550.

[25] Zhang, Y., Wang, M., Wu, Y., Tiwari, P., Li, Q., Wang, B., & Qin, J. (2025). DialogueLLM: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. Neural Networks, 192. doi:10.1016/j.neunet.2025.107901.

[26] Zhang, Y., Chen, B., Ye, H., Gao, Z., Wan, T., Lan, L., & Xu, K. (2025). Text-guided Multimodal Fusion for the Multimodal Emotion and Intent Joint Understanding. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 1–2. doi:10.1109/ICASSP49660.2025.10890680.

[27] Zhou, Z., Guo, Y., Hao, S., & Hong, R. (2025). Multi-Modal Depression Detection in Interview via Exploring Emotional Distribution Information. IEEE Transactions on Multimedia, 27, 6872 - 6883. doi:10.1109/TMM.2025.3590939.

[28] Lian, Z., Sun, H., Sun, L., Chen, K., Xu, M., Wang, K., & Tao, J. (n.d.). Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. Proceedings of the 31st ACM International Conference on Multimedia, 9610–9614. doi:10.1145/3581783.3612836.

[29] Alqurashi, F., & Ahmad, I. (2024). A data-driven multi-perspective approach to cybersecurity knowledge discovery through topic modelling. Alexandria Engineering Journal, 107, 374–389. doi:10.1016/j.aej.2024.07.044.

[30] Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation, 1(1), 67–82. doi:10.1109/4235.585893.

[31] Igel, C., & Toussaint, M. (2004). A No-Free-Lunch theorem for non-uniform distributions of target functions. Journal of Mathematical Modelling and Algorithms, 3(4), 313–322. doi:10.1023/B:JMMA.0000049381.24625.f7.

[32] Sharma, M., & Kaur, P. (2021). A Comprehensive Analysis of Nature-Inspired Meta-Heuristic Techniques for Feature Selection Problem. Archives of Computational Methods in Engineering, 28(3), 1103–1127. doi:10.1007/s11831-020-09412-6.

[33] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4), 335–359. doi:10.1007/s10579-008-9076-6.

[34] Antoniou, N., Katsamanis, A., Giannakopoulos, T., & Narayanan, S. (2023). Designing and Evaluating Speech Emotion Recognition Systems: A Reality Check Case Study with IEMOCAP. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2023-June, 1–5. doi:10.1109/ICASSP49357.2023.10096808.

[35] Zubiaga, A. (2023). Natural language processing in the era of large language models. Frontiers in Artificial Intelligence, 6, 1350306. doi:10.3389/frai.2023.1350306.

[36] Jamthe, S. (2026). Generative AI. Generative AI, 66(1), 1–10. doi:10.1201/9788743808145-1.

[37] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 5999–6009. doi:10.1201/9781003561460-19.

[38] Cuconasu, F., Trappolini, G., Siciliano, F., Filice, S., Campagnano, C., Maarek, Y., Tonellotto, N., & Silvestri, F. (2024). The Power of Noise: Redefining Retrieval for RAG Systems. SIGIR 2024 - Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 719–729. doi:10.1145/3626772.3657834.

[39] Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Alobeidli, H., Cappelli, A., ... & Launay, J. (2023). The refinedweb dataset for falcon LLM: Outperforming curated corpora with web data only. Advances in Neural Information Processing Systems, 36, 79155-79172. doi:10.5555/3666122.3669586.

[40] Chen, Z., Jiang, F., Chen, J., Wang, T., Yu, F., Chen, G., ... & Li, H. (2023). Phoenix: Democratizing ChatGPT across languages. arXiv preprint arXiv:2304.10453. doi:10.48550/arXiv.2304.10453.

[41] Liu, T., & Low, B. K. H. (2023). Goat: Fine-tuned llama outperforms gpt-4 on arithmetic tasks. arXiv Preprint, arXiv:2305.14201. doi:10.48550/arXiv.2305.14201.

[42] Hamdipour, A., Basiri, A., Zaare, M., & Mirjalili, S. (2025). BAHA: Binary artificial hummingbird algorithm for feature selection. Journal of Computational Science, 92. doi:10.1016/j.jocs.2025.102686.

[43] Hamad, R. K., & Rashid, T. A. (2024). GOOSE algorithm: a powerful optimization tool for real-world engineering challenges and beyond. Evolving Systems, 15(4), 1249–1274. doi:10.1007/s12530-023-09553-6.

[44] Suyanto, S., Ariyanto, A. A., & Ariyanto, A. F. (2022). Komodo Mlipir Algorithm. Applied Soft Computing, 114, 108043. doi:10.1016/j.asoc.2021.108043.

[45] Jiao, Y., & Du, P. (2016). Performance measures in evaluating machine learning based bioinformatics predictors for classifications. Quantitative Biology, 4(4), 320–330. doi:10.1007/s40484-016-0081-2.

[46] Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences of the United States of America, 116(44), 22071–22080. doi:10.1073/pnas.1900654116.

[47] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, 3982–3992. doi:10.18653/v1/D19-1410.

[48] Chefer, H., Gur, S., & Wolf, L. (2021). Transformer Interpretability Beyond Attention Visualization. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 782–791. doi:10.1109/CVPR46437.2021.00084.

[49] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE, 10(7), 130140. doi:10.1371/journal.pone.0130140.

[50] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. 34th International Conference on Machine Learning, ICML 2017, 7, 5109–5118.

[51] Sturm, I., Lapuschkin, S., Samek, W., & Müller, K. R. (2016). Interpretable deep neural networks for single-trial EEG classification. Journal of Neuroscience Methods, 274, 141–145. doi:10.1016/j.jneumeth.2016.10.008.

[52] Li, J. L., & Lee, C. C. (2019). Attention Learning with Retrievable Acoustic Embedding of Personality for Emotion Recognition. 2019 8th International Conference on Affective Computing and Intelligent Interaction, ACII 2019, 171–177. doi:10.1109/ACII.2019.8925536.

[53] Zou, S. H., Huang, X., Shen, X. D., & Liu, H. (2022). Improving multimodal fusion with Main Modal Transformer for emotion recognition in conversation. Knowledge-Based Systems, 258, 109978. doi:10.1016/j.knosys.2022.109978.

[54] Wang, C., Ren, Y., Zhang, N., Cui, F., & Luo, S. (n.d.). Speech emotion recognition based on multi‐feature and multi‐lingual fusion. Multimedia Tools and Applications, 81(4), 4897‐4907.

[55] Chamishka, S., Madhavi, I., Nawaratne, R., Alahakoon, D., De Silva, D., Chilamkurti, N., & Nanayakkara, V. (2022). A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling. Multimedia Tools and Applications, 81(24), 35173–35194. doi:10.1007/s11042-022-13363-4.

[56] Chauhan, K., Sharma, K. K., & Varma, T. (2023). Improved Speech Emotion Recognition Using Channel-wise Global Head Pooling (CwGHP). Circuits, Systems, and Signal Processing, 42(9), 5500–5522. doi:10.1007/s00034-023-02367-6.

[57] Singh, P., Srivastava, R., Rana, K. P. S., & Kumar, V. (2021). A multimodal hierarchical approach to speech emotion recognition from audio and text. Knowledge-Based Systems, 229, 107316. doi:10.1016/j.knosys.2021.107316.

[58] Zhang, K., Li, Y., Wang, J., Wang, Z., & Li, X. (2021). Feature fusion for multimodal emotion recognition based on deep canonical correlation analysis. IEEE Signal Processing Letters, 28, 1898–1902. doi:10.1109/LSP.2021.3112314.

[59] Yoon, Y. C. (2022). Can We Exploit All Datasets? Multimodal Emotion Recognition Using Cross-Modal Translation. IEEE Access, 10, 64516–64524. doi:10.1109/ACCESS.2022.3183587.

[60] Braunschweiler, N., Doddipatla, R., Keizer, S., & Stoyanchev, S. (2022). Factors in Emotion Recognition with Deep Learning Models Using Speech and Text on Multiple Corpora. IEEE Signal Processing Letters, 29, 722–726. doi:10.1109/LSP.2022.3151551.

[61] Abid, A., Abdalla, A., Abid, A., Khan, D., Alfozan, A., & Zou, J. (2019). Gradio: Hassle-free sharing and testing of ml models in the wild. arXiv Preprint, arXiv:1906.02569. doi:10.48550/arXiv.1906.02569.

Appendix I

Table A.1. Summary of the acronyms used

Acronyms	Descriptions
3D CNN	3-Dimensional Convolutional Neural Network
ADD	Automatic Depression Detection
AER-LLM	Ambiguity-aware Emotion Recognition Leveraging Large Language Models
AI – HUB	Artificial Intelligence Hub
AMIGOS	A Dataset for Affect, Personality and Mood Research on Individuals and Groups
ASR	Automatic Speech Recognition
BAHA	Binary Artificial Hummingbird Algorithm
Bi-GRU	Bi-Gated Recurrent Units
Bi-LSTM	Bi-Long Short-Term Memory
CCC	Concordance Correlation Coefficient
CH SIMS V2	Chinese single- and multi-modal sentiment analysis version 2
CHERMA	CHinese Emotion Recognition dataset with Modality-wise Annotions
CMU-MOSI	Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis
CNNs	Convolutional Neural Networks
DeBERTa V3	Decoding-enhanced BERT with Disentangled Attention version 3
DNN	Deep Neural Network
EAI-LLM	Emotion-Action Interpreter based on a Large Language Model
EATD	Emotional Audio-Textual Depression
E-DAIC	Extended Distress Analysis Interview Corpus
EEG	Electroencephalography
EGBM	Emotional Gestures and Body Movements
EmoryNLP	Emory Natural Language Processing
ERC	Emotion recognition in conversations
E.g.	Exempli gratia
GA	Goose Algorithm
GCM-Net	Graph-enhanced Cross-Modal Infusion with a Metaheuristic-Driven Network
GPT	Generative Pre-trained Transformer
GPT-2	Generative Pre-trained Transformer-2
GPT-3	Generative Pre-trained Transformer-3
GPT-4	Generative Pre-trained Transformer-4
GSE	General Self-Efficacy
IEMOCAP	Interactive Emotional Dyadic Motion Capture
IntervEEG-LLM	Intervention Electroencephalography - Large Language Model
ISEAR	International Survey on Emotion Antecedents and Reactions
JRBM	Joint Recognition Balance Metric
KDAE	Kinematic Dataset of Actors Expressing Emotions
KMA	Komodo Mlipir Algorithm
LFHIN	LLM Features-based Hierarchical Interaction Network
LIWC	Linguistic Inquiry and Word Count
LLM	Large Language Model
LLMs	Large Language Models
LoRA	Low-Rank Adaptation
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MC-EIU	Emotion and Intent Joint Understanding in Multimodal Conversation

MEIJU	Multimodal Emotion and Intent Joint Understanding
MELD	Multimodal Emotion Lines Dataset
MMER-LMF	Multi-Modal Emotion Recognition in Lightweight Modality Fusion
MODMA	Multi-Modal Open Dataset for Mental-disorder Analysis
MOSEI	Multimodal Opinion Sentiment and Emotion Intensity
MOSI	Multimodal Opinion-level Sentiment Intensity dataset
MSA	Multimodal Sentiment Analysis
MSE-Adapter	Multimodal Sentiment Analysis and Emotion Recognition Adapter
MSP-Podcast	Multimodal Signal Processing - Podcast
NFL	No Free Lunch
PSO	Particle Swarm Optimization
QPSO	Quantum-behaved Particle Swarm Optimization
RFS	Reliable Fusion Strategy
RLF	Reliable Learning Framework
RNN	Recurrent Neural Network
TGM	Text-Guide-Mixer
TLD	Truncated Laplace distribution
TSFL	Task-Specific Feature Learning