Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Comparative Study on Fusion Method Based on Multimodal Speech Emotion Recognition of Speech and Text

Xinjie Xie

College of Enconomy and Management, China University of Geosciences, Wuhan, China

xxj20221004430@cug.edu.cn

Abstract. With the rapid development of artificial intelligence and deep learning technologies, emotion recognition has gradually become an important research area in human-computer interaction. However, in the current gaming industry, emotion recognition is rarely utilized to optimize NPC (non-player character) intelligence to enhance immersion. Therefore, this study primarily explores the feasibility of applying multimodal emotion recognition in gaming scenarios, aiming to improve the accuracy of emotion recognition through the combination of speech and textual information, thereby optimizing NPC interactions within games. The study employs the IEMOCAP dataset, integrating audio and textual features, and conducts training and evaluation using various machine learning and deep learning models. Additionally, it compares the accuracy and training speed of several advanced fusion models to investigate whether these technologies can meet the accuracy and real-time requirements for gaming applications. The results reveal that the bimodal audio-text models significantly outperform unimodal models, with an improvement exceeding 15%. Current advanced models achieve an accuracy of over 75% with relatively short training times, preliminarily meeting the requirements for accuracy and real-time application in games.

Keywords: Emotion Recognition; Multimodal; Deep Learning; Intelligent NPC Interaction; Game.

1. Introduction

With the advancement of artificial intelligence and deep learning technologies, emotion recognition has gradually become a crucial research direction in the field of human-computer interaction and is currently applied across multiple domains [1]. Multimodal emotion recognition shows significant application potential in gaming scenarios [2], yet few mainstream games have utilized emotion recognition technology to optimize the gaming experience. In current mainstream games, traditional NPC interactions are fixed, typically developed based on hypothetical player models, allowing different interactions only through pre-selected options chosen by players. However, players' emotional states may change significantly during different game sessions, a situation that conventional games fail to accommodate [3]. Especially in open-world, RPG, and narrative-driven games, players expect NPCs to respond intelligently to their emotional changes, a demand unmet without leveraging emotion recognition technology. Therefore, integrating voice emotion recognition systems with deep learning technology is expected to enhance the intelligence of game NPCs, enabling them to adjust dialogue, behavior, and even game progression based on players' emotional states, thereby creating a more realistic and immersive gaming experience. Beyond adjusting NPC interactions, dynamically adjusting game difficulty based on emotions can significantly enhance the gaming experience. The FER dynamic balancing system developed by M. Taufik Akbar et al. adjusts in-game variables based on recognized emotions, and games using this system have significantly improved player experiences [4]. Research by Lin, W et al. also indicates that emotion-adaptive interactive games incorporating emotion recognition systems can greatly enhance player experiences [5].

This study aims to optimize intelligent interaction between NPCs and players in games. Through emotion recognition technology, NPCs can adjust dialogue, tone, and behavior in real-time according to players' emotional changes, even affecting the game narrative, making interactions more personalized. Additionally, applying emotion recognition technology allows NPCs to understand

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

"how players say" rather than just "what players say." This research will provide game developers with a new AI interaction approach, contributing to future applications of emotion recognition, especially in VR games, which can create a more immersive experience.

Currently, Ma, Ziyang et al. have compared the performance of various models across different modalities [6]. Their research indicates that combining voice and text modalities can significantly improve accuracy, while the Emotion2vec voice feature extraction model proposed by Gaurav Sahu can greatly enhance the accuracy of the voice modality [7]. The BERT model also offers substantial performance improvements in text feature extraction [8]. Furthermore, the research by Harmeen M. Saleem et al. demonstrates that deep learning models significantly enhance recognition accuracy [9]. Therefore, this study will focus on training models using these two feature extraction methods and conduct comparative analysis on different machine learning and deep learning models, exploring their feasibility in gaming applications.

2. Methods and models

2.1. Technology Roadmap

As shown in Figure 1, this is the technology roadmap for this study, covering all experimental processes and methods involved in the research.

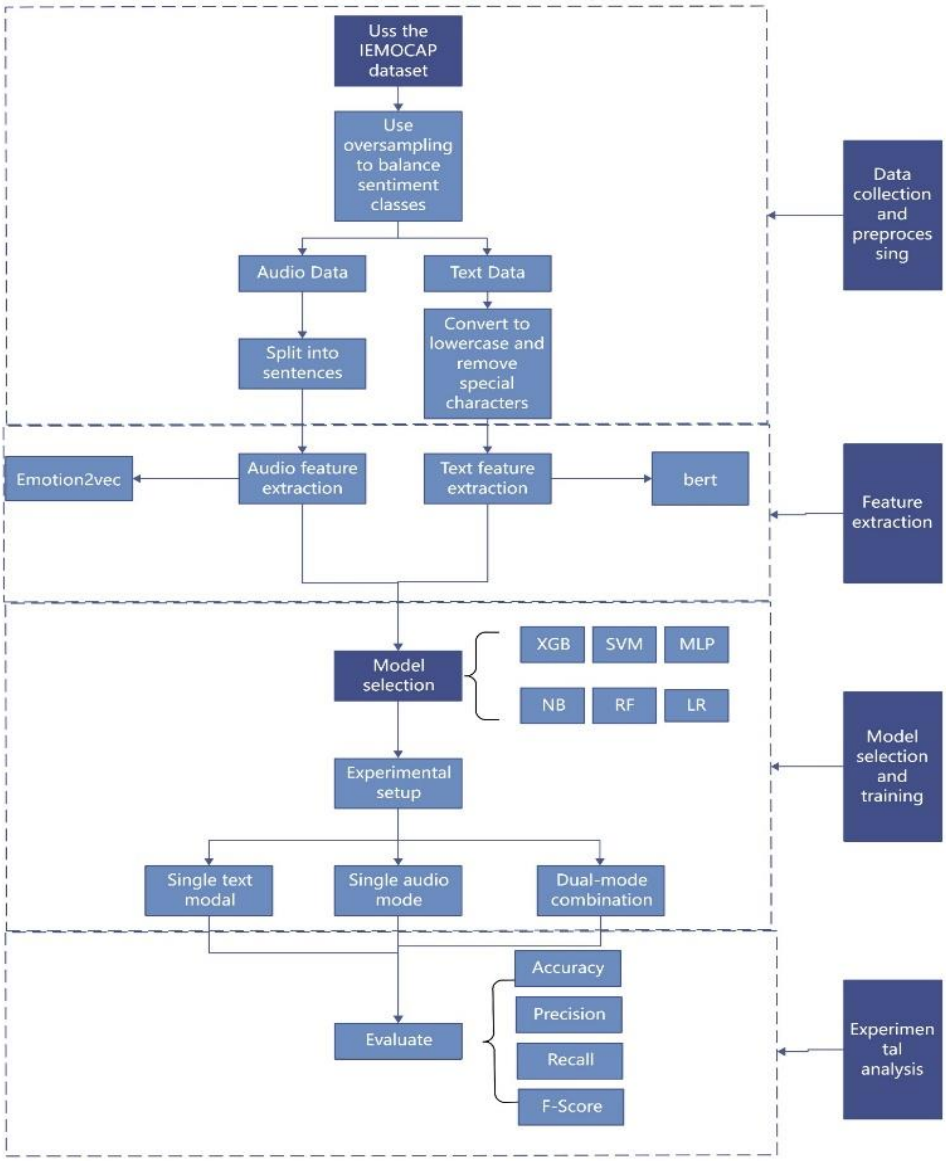

Figure 1. Technology Roadmap (Picture credit: Original)

2.2. Dataset and Preprocessing

2.2.1. Dataset Selection

The IEMOCAP dataset was selected for this study. Created by the University of Southern California, the IEMOCAP dataset is specifically designed for emotion analysis research. It includes performances from 5 male and 5 female actors, approximately 12 hours of audio, and corresponding text information, covering 8 emotion labels: anger, happiness, sadness, fear, surprise, neutral, frustration, and excitement [10]. However, in the experiment, due to the small sample size of happiness and its similarity to excitement, these two categories were merged. Ultimately, six emotion labels were chosen: anger, happiness, sadness, fear, surprise, and neutral.

2.2.2. Data Preprocessing

The data preprocessing process involved cleaning both audio and text data. For the audio data in the IEMOCAP dataset, the study first segmented the audio files in each session into sentences based on the provided transcription information. For text data, all transcriptions were converted to lowercase, and special symbols and punctuation marks were removed. Additionally, meaningless words such as "the" and "a" were excluded.

Due to the imbalance in the fear and surprise emotion categories within the dataset, oversampling techniques were applied. Specifically, SMOTE (Synthetic Minority Oversampling Technique) was used to interpolate between minority samples to balance these categories. Furthermore, samples labeled as "other" with unclear emotional annotations were removed. The final sample counts for each category are as shown Table 1

Table 1. Sample distribution

Method	Accuracy
anger	**1103**
happy	**1636**
sad	**2933**
fear	**1240**
surprise	**1177**
neu	**1708**
total	**9797**

2.2.3. Feature Extraction

Audio Features:

In this study, the Emotion2vec model was employed as the feature extractor for the speech modality. Emotion2vec is specifically designed for extracting emotional features from speech. It leverages a self-supervised learning approach and is pre-trained on large-scale speech data, enabling it to capture emotional characteristics in speech without relying on extensive labeled data. Emotion2vec has demonstrated outstanding performance in emotion analysis tasks [7]. The final audio feature dimension used in this study is 768.

Text Features:

For the text modality, the BERT model was used as the feature extractor. BERT, proposed by the Google AI Language team in 2018, is a pre-trained language representation model. It employs a bidirectional Transformer encoder architecture to achieve deep bidirectional language representation, delivering exceptional performance in natural language processing tasks [8]. The final text feature dimension used in this study is 768.

2.2.4. Feature Concatenation

To combine the two types of features, this study compared five different methods: simple_concat,

weighted_sum, attention, moe, attention_moe. After conducting simple training for each method, the results are summarized as shown in Table 2

Table 2. Concat Method Comparison

Method	Accuracy
Simple_concat	**0.7291**
Weighted_sum	**0.7133**
attention	**0.7303**
moe	**0.7082**
Attention_moe	**0.7286**

It can be observed that the best-performing methods are the attention mechanism and simple concatenation, with their metrics being almost identical. Therefore, this study applied these two methods to various models for further testing. The results showed that the overall performance of simple concatenation was approximately 5% higher. This may be because the Emotion2vec model is already a highly abstract pre-trained model, and the complex attention mechanism might disrupt its original emotional representation structure, thereby reducing performance.

2.2.5. Model Architecture

Logistic Regression (LR): Logistic regression is a linear classification model that maps the results of linear regression to probability values using the sigmoid function for classification tasks. It is simple, easy to implement, and suitable for linearly separable problems. However, it has poor fitting ability for nonlinear problems and is easily affected by correlations between features.

Random Forest (RF): Random forest is an ensemble learning method based on decision trees. It improves accuracy and reduces overfitting by constructing multiple decision trees and combining their prediction results. It can handle large-scale data and avoids overfitting, offering good robustness. However, it can be slow for real-time predictions due to its complexity.

XGBoost: XGBoost is an efficient gradient boosting algorithm that uses the Boosting method to handle large-scale data effectively and can automatically process missing values. It is efficient, powerful, and supports parallel computation, delivering excellent performance. However, it is prone to overfitting, especially when the data contains significant noise.

Support Vector Machine (SVM): SVM is a supervised learning model mainly used for classification tasks. It achieves classification by finding the optimal separating hyperplane in high-dimensional space. This model performs well, especially for high-dimensional data, and has good performance even with small sample sizes. However, it is slow for training on large datasets and is sensitive to the choice of kernel functions.

Multilayer Perceptron (MLP): MLP is a feedforward neural network that performs nonlinear transformations through multiple hidden layers, making it suitable for complex classification tasks. It can handle nonlinear problems and is applicable to large-scale data. However, it requires long training times, is prone to local optima, and demands large amounts of data.

BiLSTM: This study employs a bidirectional long short-term memory network (BiLSTM) as one of the training models. BiLSTM, first proposed by Graves and Schmidhuber in 2005, is an extension of traditional LSTM. By processing sequential data in both forward and backward directions, it captures richer contextual information. The hidden states from both directions are typically combined through concatenation or summation to form the final feature representation. This bidirectional processing mechanism makes BiLSTM particularly suitable for natural language processing tasks that require consideration of complete context.

3. Experiment and Results

3.1. Experimental Setup

This study's experimental setup involves three modalities for comparative research: Audio Modality, Text Modality, and Audio+Text Modality. Evaluation metrics include Accuracy, Precision, Recall, and F-score.

During training, the number of training iterations is uniformly set to 100, and an early stopping mechanism is employed to prevent overfitting. Early stopping is triggered when the loss rate does not improve for more than 10 consecutive iterations.

3.2. Experimental Results

3.2.1. Text Modality

As shown in Table 3, the optimal model SVC performs best across all metrics, achieving an accuracy of 69.9%.

However, using the text modality alone has obvious limitations, such as the inability to recognize sarcasm, Text ambiguity. These issues result in accuracy consistently below 70%, performing worse than the audio modality. From the confusion matrix, it can be observed that sur (surprise) and ang (anger) are recognized exceptionally well; hap (happiness), sad (sadness), fea (fear), and neu (neutral) emotions show significant confusion. This indicates that distinguishing these emotions is challenging in pure text expressions. Figure 2 shows the Text Modularity Confusion Matrix

Table 3. Text Modality Performance

Method	Accuracy	F-score	Precision	Recall
RF	**0.6597**	**0.6453**	**0.6737**	**0.6597**
XGB	**0.676**	**0.6696**	**0.6767**	**0.676**
SVC	**0.699**	**0.6927**	**0.7012**	**0.699**
MLP	**0.6745**	**0.671**	**0.6688**	**0.6745**
BiLSTM	**0.6724**	**0.6684**	**0.6751**	**0.6724**
LR	**0.65**	**0.6458**	**0.6432**	**0.65**
ADVERAGE	**0.702**	**0.6991**	**0.702**	**0.6963**

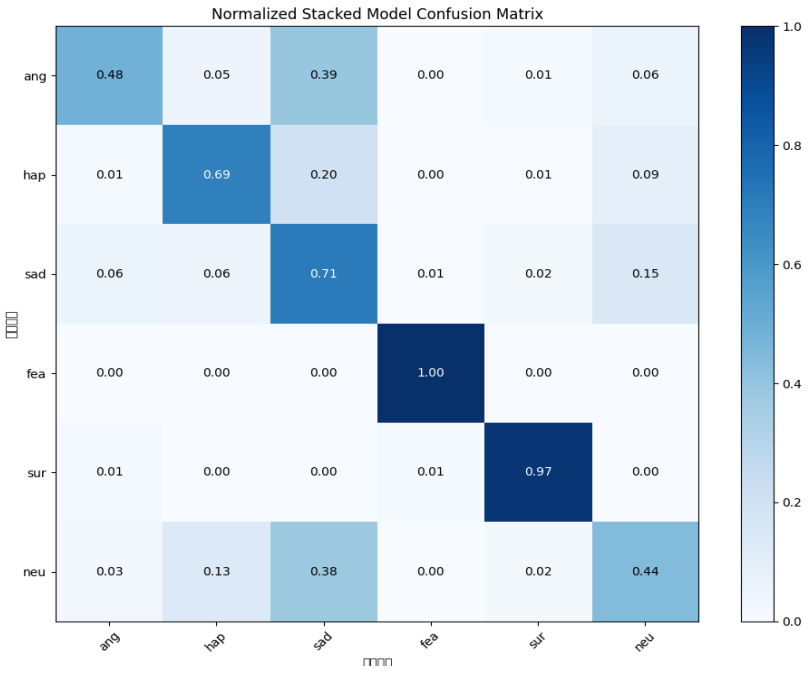

Figure 2. Text Modality Confusion Matrix (Picture credit: Original)

3.2.2. Audio Modality

As shown Table 4, when using only audio features, thanks to the powerful Emotion2vec model, the overall performance surpasses that of the text modality. Certain trained models even achieve an accuracy of over 70%. From the confusion matrix, it can be observed that fea (fear) and sur (surprise) are recognized well. The advantage of the audio modality lies in its ability to capture prosodic features in speech, such as pitch, speech rate, and energy, which play a crucial role in emotion analysis.

Table 4. Audio Modality Performance

Method	Accuracy	F-score	Precision	Recall
RF	**0.6393**	**0.6079**	**0.6717**	**0.6393**
XGB	**0.6878**	**0.6806**	**0.6931**	**0.6878**
SVC	**0.7395**	**0.7394**	**0.74**	**0.7403**
MLP	**0.7082**	**0.7056**	**0.7082**	**0.7066**
LR	**0.6791**	**0.6766**	**0.6748**	**0.6791**
BiLSTM	**0.6602**	**0.6541**	**0.6574**	**0.6602**
ADVERAGE	**0.7433**	**0.71**	**0.72**	**0.71**

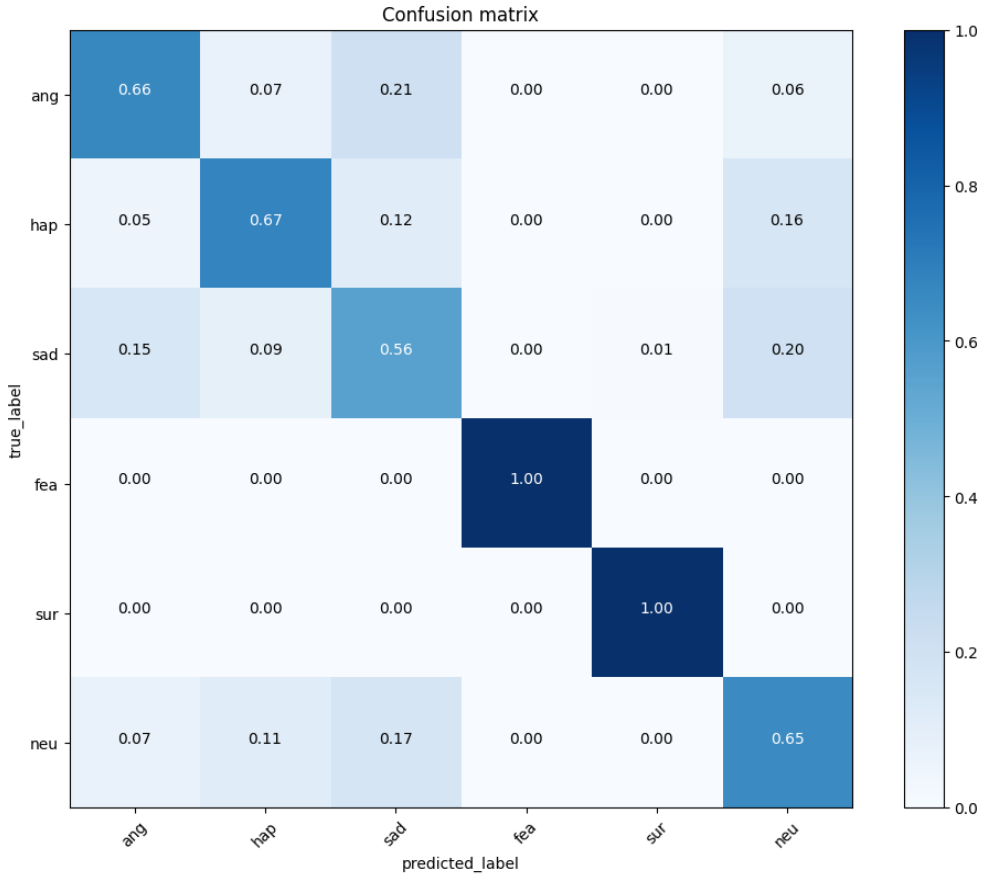

Figure 3. Audio Modality Confusion Matrix (Picture credit: Original)

3.2.3. Fusion Modality

From the initial confusion matrix as Figure 3, it is evident that the recognition performance for ang (anger) was not satisfactory. However, the audio modality demonstrated excellent recognition for ang. The inclusion of the text modality diluted the audio modality's recognition effectiveness for ang, which did not align with the expectation of complementary contributions from both modalities.

To address this, during feature concatenation, this study increased the weight of audio features for the ang label, setting the coefficient to 2.0, while keeping the coefficient for text features at 1.0. The final results are as shown Table 5.

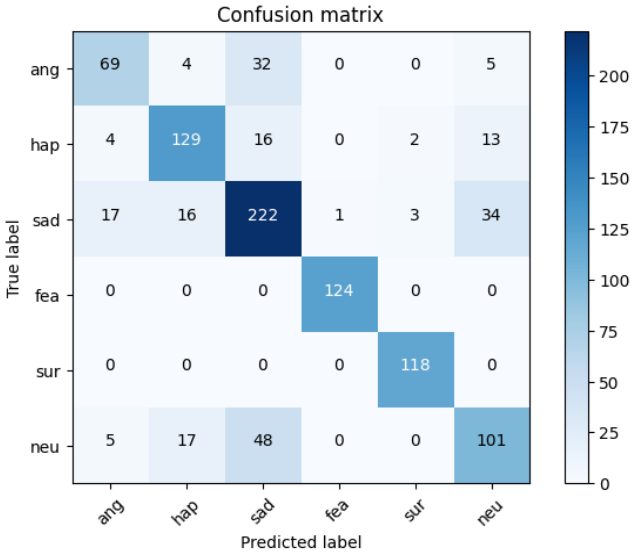

Figure 4. Initial Fusion Modality Confusion Matrix (Picture credit: Original)

Table 5. Final Fusion Modality Performance

Method	Accuracy	F-score	Precision	Recall
RF	**0.7372**	**0.7551**	**0.7985**	**0.7537**
XGB	**0.792**	**0.821**	**0.834**	**0.817**
SVC	**0.8235**	**0.8526**	**0.8525**	**0.8538**
BiLSTM	**0.749**	**0.751**	**0.748**	**0.752**
MLP	**0.779**	**0.789**	**0.797**	**0.784**
LR	**0.698**	**0.731**	**0.731**	**0.733**
ADVERAGE	**0.8439**	**0.8722**	**0.8431**	**0.8429**

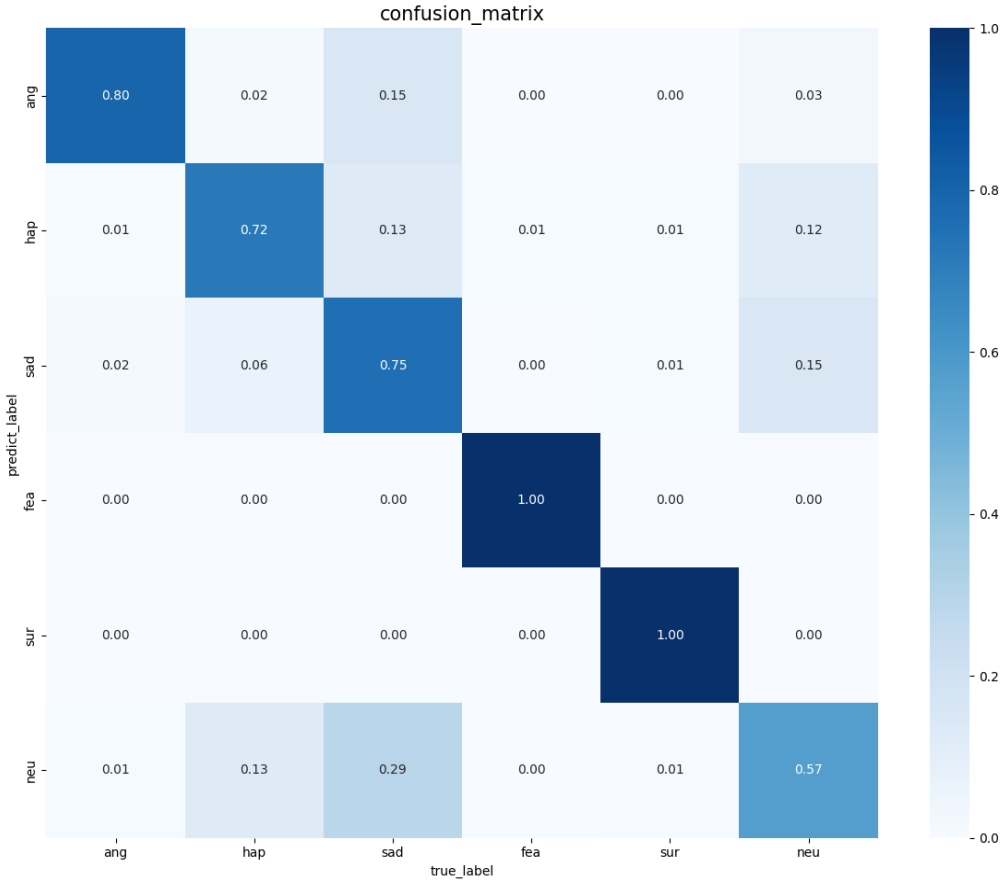

Figure 5. Final Fusion Modality Confusion Matrix (Picture credit: Original)

Figure 4 shows the Initial Fusion Morphology Confusion Matrix and Figure 5 shows the final fusion mode confusion matrix. It can be observed that after optimizing the concatenation weights, all metrics have significantly improved. From the confusion matrix, the increase in accuracy is mainly attributed to the improved recognition ability for ang (anger). The final model only shows notable confusion in recognizing hap (happiness) and neu (neutral). As the ultimate fusion modality, its metrics comprehensively outperform those of single modalities. The best-performing model in terms of accuracy is SVC, and considering real-time requirements, SVC demonstrates strong potential in balancing speed and performance. This makes it suitable for scenarios like gaming, where players' voices may need to be recorded and trained in real time.

3.2.4. Comparative Analysis

Table 6. Comparison of Experimental Results

	Text	Audio	Text+Audio
Accuracy	**0.699**	**0.7395**	**0.851**
F-score	**0.6927**	**0.7394**	**0.8786**
Precision	**0.7012**	**0.74**	**0.85**
Recall	**0.699**	**0.7403**	**0.85**

From Table 6, they provided experimental results, we can clearly observe the performance differences among the single text modality, single audio modality, and text+audio fusion modality in the emotion classification task. The performance of the three modalities exhibits a clear gradient difference: Fusion Modality > Single Audio Modality > Single Text Modality.

Performance Metrics Comparison:

Accuracy: The fusion modality achieves 0.851, representing an improvement of approximately 15.1% and 21.7% compared to audio modality (0.7395) and text modality (0.699), respectively.

F-score: The fusion modality achieves 0.8786, with improvements of 18.8% and 27.3% over audio (0.7394) and text (0.6927), respectively.
Precision: The fusion modality achieves 0.85, showing improvements of 14.8% and 21.2% over audio (0.74) and text (0.7012), respectively.

Recall: The fusion modality achieves 0.85, representing improvements of 14.8% and 21.4% over audio (0.7403) and text (0.699), respectively.
These results clearly demonstrate the complementary nature of text and audio features. Text primarily captures semantic information, while audio extracts acoustic features such as pitch, speech rate, and emotional intensity. Their combination enables a more comprehensive analysis of emotional expressions. Moreover, the inclusion of audio helps recognize special emotional expressions like sarcasm, significantly improving the accuracy of emotion recognition, especially for those emotion categories that are difficult to distinguish in single modalities. This makes it particularly suitable for analyzing complex player emotions in gaming scenarios.

A comparison of the confusion matrices for the three modalities reveals significant differences in the recognition of emotional categories. In the single text modality, only sad, fear, happy, and surprise are recognized well. However, anger and neutral show significant confusion, with poor performance. In the single audio modality, the anger category performs significantly better than the text modality. While the recognition rate for happy is slightly lower than the text modality, it still performs well. The sad category performs poorly, while sad and surprise are exceptionally well-recognized. The neutral category performs moderately but better than the text modality. The addition of the audio modality significantly improves emotion recognition performance, particularly by reducing the confusion between anger and other categories. The recognition rate for anger increased significantly to 0.76, while other emotions also showed improvements. Furthermore, the already excellent recognition rates for fear and surprise remained at 100% accuracy. This highlights the imbalance in emotion category recognition: fear and surprise are easy to recognize across all modalities, while

neutral is more challenging to distinguish. The fusion modality significantly improves the recognition performance for anger, happiness, and sadness, but provides limited improvement for neutral.

By comparing the confusion matrices of different modalities, we can observe significant modality complementarity. The text modality outperforms the audio modality in recognizing sad emotions. The audio modality excels in recognizing neutral and anger emotions compared to the text modality. In the text modality, anger and neutral are frequently misclassified as sad, whereas in the audio modality, this error rate is significantly reduced. In the fusion modality, these errors are notably complemented, with a significant reduction in misclassified samples, leading to improved accuracy. Additionally, in the fusion modality, sad being misclassified as neutral is significantly lower than in the single text or audio modalities. This indicates that text and audio modalities capture multidimensional aspects of emotional expression. Combining these complementary features can significantly enhance the accuracy of emotion recognition. For anger and sad, in particular, the fusion modality proves to be especially effective.

Different models exhibit distinct performance when handling multi-modal data. SVC performs the best across all three modalities, with the greatest improvement in fusion performance. From single audio to fusion modality, SVC achieves an 11.3% improvement. Its strength lies in its ability to handle high-dimensional feature spaces and nonlinear decision boundaries, making it particularly effective in capturing complex feature interactions in emotion analysis. This makes it stand out in multi-modal recognition tasks. XGBoost ranks second across all modalities, with a 15.2% improvement from single audio to fusion modality. Its ensemble tree model effectively captures nonlinear relationships and interaction effects among features, performing exceptionally well when handling heterogeneous concatenated features. Although the performance of BiLSTM and MLP is not as strong as traditional machine learning methods like SVC, this is primarily due to the limited dataset size (fewer than 10,000 samples). MLP shows impressive performance, achieving a 10% improvement from single audio to fusion modality. Its accuracy is second only to SVC and XGBoost, as its network structure is well-suited for learning complex nonlinear relationships between modalities. While BiLSTM's accuracy is not high, it achieves a significant improvement of 13.4%, highlighting its unique advantage in capturing temporal features in sequential data for emotional expression. Deep learning models exhibit great potential, and increasing the training dataset size may allow them to surpass SVC in accuracy. These results indicate that for medium-scale multi-modal emotion analysis, SVC (kernel-based method) and XGBoost (tree-based ensemble method) are the most effective algorithm choices. However, for larger-scale datasets, selecting deep learning models is likely to yield better results.

4. Application Discussion for Games

Games are characterized by their high degree of human-computer interactivity, offering players emotional experiences that are far more complex than those provided by traditional media, while also demanding a higher level of immersion. The experimental results of this study demonstrate that multimodal emotion analysis, combining text and audio, significantly improves the accuracy of emotion recognition. This effectively addresses the issue of low accuracy in emotion recognition when relying solely on audio, thereby enhancing emotional interaction within games. Game environments typically involve diverse and complex information sources, such as player voice commands, text-based chats, and in-game audio feedback. Multimodal emotion analysis can integrate emotional information from these various types of data, enabling games to accurately interpret players' emotional states and deliver more natural and immersive gaming experiences. For instance, in RPG games, pre-selected options can be hidden, allowing NPCs to interact with players and infer their current emotional state through dialogue. This enables NPCs to automatically select the most suitable response, providing a low-cost implementation of this technology for intelligent NPC interactions.

The real-time nature of gaming imposes strict requirements on the speed of emotion analysis. According to the experimental results, the SVC model performs best in the multimodal fusion setting, with a computational complexity of (O(n^2d)) and a recognition time of 30–45ms. The XGB model achieves comparable metrics to SVC but has a significantly lower computational complexity of (O(KD)), striking a good balance between processing speed and accuracy. Tests indicate that under current hardware conditions, the XGB model can complete recognition within 20–30ms. This demonstrates the feasibility of real-time emotion analysis in games under current technological constraints. Particularly for conversational games, RPGs, and narrative-driven games, which have relatively lenient latency requirements, existing multimodal emotion analysis techniques are already capable of meeting basic demands. Furthermore, the experiments reveal that single-audio modality outperforms single-text modality, offering a practical solution for performance-constrained platforms, such as mobile devices or older hardware. In such cases, prioritizing audio-based emotion analysis can serve as an effective compromise.

Different types of games require the recognition of varying emotional states. The findings of this study indicate that multimodal emotion recognition struggles primarily with neutral emotions, which have lower recognition accuracy. However, in practical gaming applications, neutral emotions are often not critical for emotion recognition. Thus, the current level of accuracy is sufficient to enable precise NPC interactions. For example, in horror games, the primary emotions to be recognized are fear and surprise. Multimodal emotion recognition achieves exceptionally high accuracy for these emotions, reaching 100%. Even single-modality systems can accurately capture players' fear responses, enabling highly responsive horror experiences that dynamically adjust game elements based on the intensity of players' fear.

Based on the experimental results and current trends in the gaming industry, the future development of multimodal emotion analysis in games will likely involve the integration of additional modalities. Beyond text and audio, facial expressions and physiological signals could be incorporated, especially in VR games, where VR devices can facilitate the collection of various physiological signals. Systems capable of storing and learning players' emotional data could also be developed to continuously refine and optimize emotion recognition models as the game progresses. Additionally, advanced large language models could be integrated to generate personalized narratives and content in real-time based on players' emotions and dialogue. These innovative directions highlight the immense potential of multimodal emotion analysis in gaming applications. Our experimental results confirm that multimodal methods combining text and audio significantly enhance the accuracy of emotion recognition, providing a solid technical foundation for emotional interaction systems in games. While challenges such as privacy concerns and environmental noise remain in real-world gaming scenarios, these issues are not insurmountable. As technology continues to advance and the gaming industry's demand for affective computing grows, multimodal emotion analysis is poised to become a core technology in future game design, enabling games to truly understand and respond to players, thereby delivering personalized gaming experiences for everyone.

5. Conclusion

Comparative experiments have demonstrated that the feature extraction methods Emotion2vec and BERT exhibit excellent performance. Combining features extracted by these two methods for multimodal fusion significantly enhances emotion recognition performance, achieving an accuracy of 85.1%, which represents a 15.1% improvement compared to single-modality audio and a 21.7% improvement compared to single-modality text. The audio modality excels in recognizing tone-sensitive emotions (e.g., anger and neutral), while the text modality is more adept at capturing semantic-related emotions (e.g., sadness and fear). The complementary nature of these modalities provides a new paradigm for analyzing complex emotions in gaming scenarios.

The experiments further reveal that traditional machine learning models (SVC and XGBoost) outperform deep learning models (BiLSTM and MLP) when working with medium-sized datasets.

Among these, XGBoost achieves the optimal balance between accuracy and inference speed, making it well-suited for real-time interactive gaming requirements. This technology enhances immersion through dynamic NPC responses, particularly meeting the intelligent interaction needs of horror and RPG games. It offers a scalable framework for the application of affective computing in the gaming domain.

References

[1] N. Ahmed, Z. A. Aghbari, S. Girija, A systematic survey on multimodal emotion recognition using learning algorithms, Intelligent Systems with Applications, Volume 17, (2023), 200171, ISSN 2667-3053.

[2] K. K. Smith, B. Victoria, S. Esmaeil. Nadimi, U. Rajendra Acharya, Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations, Information Fusion, Volume 102, (2024),102019, ISSN 1566-2535.

[3] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, M. R. Wróbel. Emotion Recognition and Its Applications. In: Hippe, Z., Kulikowski, J., Mroczek, T., Wtorek, J. (eds) Human-Computer Systems Interaction: Backgrounds and Applications 3. Advances in Intelligent Systems and Computing, vol 300. Springer, Cham. (2014).

[4] M. Taufik Akbar, M. Nasrul Ilmi, V. Imanuel. J. Moniaga, T. K. Chen, A. Chowanda. Enhancing Game Experience with Facial Expression Recognition as Dynamic Balancing, Procedia Computer Science, Volume 157, (2019), Pages 388-395, ISSN 1877-0509.

[5] W. Lin, C. Li, Y. Zhang. A System of Emotion Recognition and Judgment and Its Application in Adaptive Interactive Game. Sensors (2023), 23, 3250.

[6] Z. Y. Ma, Z. S. Zheng, J. X. Ye, J. C. Li, Z. F. Gao, S. L. Zhang, X. Chen, Multimodal Speech Emotion Recognition and Ambiguity Resolution, arXiv:1904.06022 [cs. LG], (2024).

[7] G. Sahu, Emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation, Proc. ACL 2024 Findings, 12 Apr (2019).

[8] D. Jacob and M. Wei and K. Lee and T. Kristina, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805, (2018).

[9] M. Sharmeen, S. A. Abdullah, Y. Siddeeq. A. Ameen, A. M. Mohammed, S. Zeebaree. Multimodal Emotion Recognition Using Deep Learning. JASTT (2021), 2 (01), 73-79.

[10] C. Busso, M. Bulut, C. Lee, et al. IEMOCAP: interactive emotional dyadic motion capture database. Lang Resources & Evaluation 42, 335–359 (2008).