frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 30 January 2024
DOI 10.3389/fnbot.2023.1324831

Multimodal robotic music performance art based on GRU-GoogLeNet model fusing audiovisual perception

Lu Wang*

School of Preschool and Art Education, Xinyang Vocational and Technical College, Xinyang, China

Check for updates

OPEN ACCESS

EDITED BY
Mehdi Benallegue,
UMI3218 Joint Robotics Laboratory (JRL),
Japan

REVIEWED BY
Wei Zhou,
University of Waterloo, Canada
Jianchu Lin,
Huaiyin Institute of Technology, China
Sahraoui Dhelim,
University College Dublin, Ireland

*CORRESPONDENCE
Lu Wang
✉ wangluzy@163.com

RECEIVED 20 October 2023
ACCEPTED 06 December 2023
PUBLISHED 30 January 2024

CITATION
Wang L (2024) Multimodal robotic music performance art based on GRU-GoogLeNet model fusing audiovisual perception.
Front. Neurorobot. 17:1324831.
doi: 10.3389/fnbot.2023.1324831

The field of multimodal robotic musical performing arts has garnered significant interest due to its innovative potential. Conventional robots face limitations in understanding emotions and artistic expression in musical performances. Therefore, this paper explores the application of multimodal robots that integrate visual and auditory perception to enhance the quality and artistic expression in music performance. Our approach involves integrating GRU (Gated Recurrent Unit) and GoogLeNet models for sentiment analysis. The GRU model processes audio data and captures the temporal dynamics of musical elements, including long-term dependencies, to extract emotional information. The GoogLeNet model excels in image processing, extracting complex visual details and aesthetic features. This synergy deepens the understanding of musical and visual elements, aiming to produce more emotionally resonant and interactive robot performances. Experimental results demonstrate the effectiveness of our approach, showing significant improvements in music performance by multimodal robots. These robots, equipped with our method, deliver high-quality, artistic performances that effectively evoke emotional engagement from the audience. Multimodal robots that merge audio-visual perception in music performance enrich the art form and offer diverse human-machine interactions. This research demonstrates the potential of multimodal robots in music performance, promoting the integration of technology and art. It opens new realms in performing arts and human-robot interactions, offering a unique and innovative experience. Our findings provide valuable insights for the development of multimodal robots in the performing arts sector.

KEYWORDS

emotion, music, multimodal robot, GRU, GoogLeNet

1 Introduction

With the rapid development of technology, multimodal robots are becoming more and more common in real life, covering a number of different fields, and have been widely used in all aspects of daily life. They are deployed in healthcare, education, customer service, manufacturing, entertainment, autonomous vehicles, agriculture, search and rescue, home assistance, environmental monitoring, retail, inventory management, and public safety (Erickson et al., 2020). These robots enhance patient care, assist in surgeries, provide interactive education, improve customer experiences, automate industrial processes, entertain in theme parks, revolutionize transportation, aid in agriculture tasks, support search and rescue missions, assist with household chores and security, monitor environmental conditions, optimize retail operations, ensure public safety, and much more. The versatility of multimodal robots enables them to adapt to different tasks and environments, making them valuable tools in enhancing efficiency, safety, and overall quality of life.

Not only that, multimodal robots are being used more and more in the performing arts. Robot performing art refers to the combination of robot technology and performing art to create a unique and amazing art form (Inamura and Mizuchi, 2021). It uses robots as the main body or participants of the performance, and through the robot's movements, music, lighting, images and other elements, it presents the audience with a visual feast that integrates technology and art. Robot performance art has a variety of forms and expressions, such as dance performance, music performance, theater performance, interactive performance, and so on Wang et al. (2021). It is the intersection of technology and art, which is developing rapidly and has made significant contributions in a number of fields in recent years.

To advance the development of robotic music performance art and bring it closer to everyday life, this paper explores the novel integration of the GRU-GoogLeNet model in the field of multimodal robotic music performance art, aiming to enrich the audio-visual experience. Our study defines audiovisual perception as an integrated approach to information processing (Li et al., 2023), simultaneously parsing and synthesizing the acoustic (audio) and visual elements of a musical performance. This encompasses not just the aural understanding of the music, such as rhythm, melody, and tonality, but also the interpretation of visual information from performers and audiences (Tsiourti et al., 2019), like facial expressions, body language, and scene interaction.

This research not only fuses robotics, music, and visual arts into a unique performance style but also offers a novel artistic experience to audiences. Our methodology marks a significant advance in both art and technology, potentially influencing society and culture. During performances, robots equipped with our model can analyze the audience's emotions and reactions, dynamically adjusting the performance in real-time to establish a more profound connection with the audience. This enhances the appeal and engagement of the performing arts, elevating the quality, and immersive experience of the artwork.

Our research focuses on enhancing the integration of visual and auditory perception in robotic performances, setting a new benchmark in the field. Differing from traditional approaches, our model combines the strengths of GRU for audio data analysis and GoogleNet for visual data processing. This integration allows for a more nuanced understanding of musical performances, offering a richer and more engaging experience to the audience. Our study incorporates the latest technological advancements, aligns with contemporary research, and provides a fresh perspective on integrating advanced machine learning techniques in artistic performances. Clearly, this research represents the cutting edge of integrating art and technology, embodying significant implications. The development of this art form will help drive the joint progress of technology and art, while bringing more entertainment options to the public.

Here are some models that related to this area.

1.1 LSTM-based sentiment analysis model

To integrate audiovisual perception in multimodal robotic musical performing arts, it is important to perform sentiment analysis of musical information and user feedback. LSTM (Long Short-Term Memory) model is a type of Recurrent Neural Network (RNN), which is one of the most powerful dynamic classifiers publicly known (Staudemeyer and Morris, 2019) and well-suited for continuous data analysis. LSTM can be used for sentiment analysis to determine the emotion or mood expressed in a given text, and for analyzing music, including extracting notes, melodies, and emotions from the audio (Laghrissi et al., 2021). LSTM can also be used to analyze music, including extracting notes, melodies, and emotions from audio. However, the LSTM model has some limitations. One of the major limitations is that it struggles with long sequences. Longer text sequences require significant computational resources and time for training and inference. Capturing long-term dependencies becomes more complex as the length of the sequence increases. In addition, LSTM models for sentiment analysis and other natural language processing tasks usually require a large amount of labeled data for training to obtain accurate results (Zhou et al., 2019). However, obtaining a sufficient amount of labeled data, especially for certain languages or specific domains, can be a challenging task. This also makes training LSTM models more difficult.

1.2 Face recognition model

Face recognition modeling is a technique used to identify and verify faces and is widely used in the field of computer vision. Face recognition in a broad sense includes related technologies for building a face recognition system. It includes face detection, face position, identity recognition, image preprocessing, etc. (Li et al., 2020). Face recognition models determine identity by extracting face features and comparing them with known face features. This model usually use deep learning algorithms such as convolutional neural networks (CNN) or face embedding techniques to build efficient face recognition systems by learning a large number of face images. It is widely used in various scenarios such as security surveillance, face unlocking and face payment (Hariri, 2022). Multimodal Robotic Music Performing Art with Integrated Audiovisual Perception Music performing art is an art form that combines visual, auditory, and robotic technologies. In this kind of performing art, robots use sensors such as cameras and microphones to sense the audience's audio-visual behavior and the ambient music, and based on this information, they display movements and expressions that are coordinated with the music. By integrating audiovisual perception, robots are able to interact with the audience and present a more vivid and expressive artistic performance. However, the use of facial recognition models may also raise privacy concerns, especially as the recognition and tracking of an individual's facial information without explicit authorization may violate an individual's right to privacy. Therefore, the question of how to ensure the legal and ethical use of facial recognition technology is a crucial one (Bhat and Jain, 2023).

1.3 Natural language processing model

Natural Language Processing (NLP) modeling plays an important role in the art of multimodal robotic musical performances that integrate audiovisual perception (Qiu et al., 2020). It utilizes speech recognition technology to realize voice interaction between the robot and the audience, accurately understand the audience's words and commands, and respond and interact accordingly. At the same time, through the analysis of natural language processing models, the robot can also recognize the audience's language expression and intonation, from which it can understand the audience's mood and emotional state. This process helps the robot adjust its performance or response to better interact with the audience and create a more personalized interactive experience (Zhu et al., 2019). However, it is worth noting that there are still some challenges and shortcomings of natural language processing models in this artistic field (Li, 2022). For example, the accuracy of speech recognition can be problematic, and sometimes audience commands are misinterpreted or incorrect responses are generated. And due to the creative limitations of the model, it may result in the robot's responses being too stilted or predictable to generate responses that surprise the audience.

1.4 Interactive strategy model

The interactive strategy model has multiple perceptual capabilities and can perceive various states of the audience and the environment, including elements such as music tempo, audience's emotional feedback and body movements (Andersson et al., 2020; Pang et al., 2020). Through in-depth analysis of these perceptual information, the model can intelligently formulate corresponding interaction strategies according to different contexts, providing highly personalized performance experiences for the audience. The model can also flexibly adjust the form and content of the performance according to the audience's emotional feedback, such as voice expression and facial expression, to guide the audience to a more positive and deeper emotional experience (Lee et al., 2020). However, the interactive strategy model still has some accuracy difficulties in perceiving and understanding the audience's emotions and movements. Although existing technologies are able to perceive and analyze to a certain extent, the models may be limited in accurately understanding and responding to audience needs when facing complex music performance scenes and diverse audience behaviors. This issue requires continued research and technological improvements to improve the accuracy and adaptability of the models to better meet audience expectations and provide superior music performance experiences.

After comprehensively considering the advantages and disadvantages of other models, we have designed an innovative multimodal robotic music performance art system that integrates GRU and Google Net models. This system utilizes deep learning and artificial intelligence technologies to perceive, analyze, and generate music, and combines it with rich visual information to provide an all-encompassing artistic experience for the audience, resulting in a richer and more integrated performance experience. This system not only contains music generation and analysis functions, but also covers the ability to analyze music emotion. Through advanced emotion analysis technology, the robotic music performer is able to accurately perceive and understand the audience's emotional state, and adjust the performance form and content accordingly. As an example, when the audience expresses joyful or excited emotions, the robot performer can choose upbeat, high-energy music as well as exciting dance movements, thus further stimulating the audience's positive emotions and in-depth experience. In a multimodal robotic musical performing arts system that integrates audiovisual perception, the robot not only analyzes and generates music through music and sound perception, but also acquires information about the stage, the audience, and the performer through visual perception. This visual information includes the audience's gestures, facial expressions and movements, as well as the stage lighting, camera inputs, etc. By combining GRU and Google Net models, the system is able to deeply analyze and understand this visual information in order to better adapt to the audience's emotional state and stage effects, and further enhance the artistry and interactivity of the performance.

The contribution points of this paper are as follows:

- This multimodal robotic music performance art system utilizes advanced deep learning and artificial intelligence technologies to perceive, analyze, and generate music by fusing GRU and Google Net models and combining them with rich visual information to provide an all-encompassing art experience.
- The system is able to accurately perceive and understand the audience's emotional state through music emotion analysis technology, thus adjusting the performance form and content in real time, increasing the audience's sense of participation and immersion, enhancing the charm and interactivity of the music performance art, and creating a more interactive and personalized music performance experience for the audience.
- It applies advanced artificial intelligence technology in music analysis, emotion recognition and image processing. Through deep learning methods, the system is able to learn and extract features of music and images from a large amount of data, enabling the generation of music and recognition of emotions. This application of technology enables the robot performer to have a higher level of interpretive and creative ability, thus providing a more artistic and emotionally resonant musical performance.

2 Related work

2.1 CNN-based emotion recognition model

Emotion recognition is an important field in the performing arts of multimodal robots, and its goal is to automatically recognize the emotional state of robots by integrating multiple perceptual modalities such as vision, hearing, and touch. Research in this area has made significant progress, and one representative model is the deep learning model.

Selvi et al. proposed a deep learning-based multimodal emotion recognition model that utilizes both visual and acoustic information (Selvi and Vijayakumaran, 2023). Specifically, they used a convolutional neural network (CNN) for image information and a recurrent neural network (RNN) for sound information. These neural networks are able to extract emotional features from images and sounds, respectively, and then fuse them together to ultimately output the emotional state of the robot. However, despite the success of this approach in multimodal emotion recognition, there are still some shortcomings.

First, the accuracy of the model still needs to be improved for complex emotional states. For example, the performance of the model may degrade when recognizing mixed emotions (e.g., being both happy and anxious). Second, the generalization ability of this method is limited and it is difficult to adapt to the differences in emotional expressions caused by different cultural backgrounds and individual differences. However, the performance in multimodal tasks is somewhat limited. Traditional CNN models usually require multiple independent networks in multimodal tasks, resulting in inadequate information fusion and difficulty in adapting to different types of multimodal input data, which leads to their limited generalization for processing multimodal tasks (Pepino et al., 2021). Therefore, further research and improvements are needed to address these issues in future studies.

2.2 Multimodal fusion approach to processing information

Multimodal fusion methods are a class of methods proposed to solve the problem of information integration in multimodal data analysis, which provide a richer, more interactive and perceptual performance experience in the art of robotic music performance, while enhancing the connection and communication between the robot and the audience, making the music performance more vivid and memorable. Among them, multimodal attention mechanisms (Ghaleb et al., 2020) and multimodal convolutional neural networks (FusionNet) (Vakalopoulou et al., 2019) are two common approaches.

The multimodal attention mechanism allows the model to dynamically attend to information from different modalities in order to increase the weight of specific modalities in the task. FusionNet, on the other hand, fuses the features of multiple modalities through a specific network structure to achieve the integration of multimodal information. However, multimodal attention mechanisms usually require a large amount of labeled data for training, which is difficult to obtain in some domains, limiting their feasibility in practical applications. In addition, although methods such as FusionNet provide a way to integrate multimodal information, they usually perform well on specific tasks, lack generalizability, and are difficult to apply to different multimodal tasks.

2.3 Augmented reality modeling in the performing arts

Multimodal robotic music performance technology fuses art and technology, integrating music, robotics, and multimodal perception, bringing a new audio-visual experience to music performances. In order to enrich the performance effect, Augmented Reality (AR) Modeling has been created and combined with music performance (Petrović, 2020). This technological innovation utilizes AR technology to enable virtual elements to interact with the actual scene, providing the audience with a richer and more unique visual enjoyment. AR modeling relies on multi-disciplinary knowledge such as computer vision and sensing technology, and is able to capture audio signals in real time and present them in a virtual visual effect.

At the same time, the application of augmented reality models in multimodal music performances has some limitations. First, organically integrating multiple models may face the challenges of model interface inconsistency and performance matching, which require clever design and optimization to fully utilize the performance of the models. Moreover, music performances usually require real-time performance, so the integration of all models must maintain low latency while maintaining high performance, which increases the stringent requirements on the performance environment, as factors such as the external environment and equipment failures may adversely affect the models. In addition the application of AR models also requires appropriate hardware equipment and technical support, which may increase the cost and technical complexity (Lee, 2019).

To summarize, the application of augmented reality models in multimodal music performances brings innovation and interactivity to performances, but there are still some challenges in terms of technology and cost. However, these challenges are expected to be gradually overcome as technology continues to advance, making AR a more common and compelling element of musical performances.

3 Methods

3.1 Overview of our network

Our approach integrates GRU and GoogleNet models to enhance robotic music performing arts systems, created a model that can accept multiple modal input data, such as audio and image information. Audio information can be acquired through microphones or audio sensors, while image information can be acquired through cameras or visual sensors. These two input modalities can capture the musical and visual elements of the performing arts to provide a more comprehensive perception for the robot. The combination of these modal data can provide richer information for the creative and perceptual aspects of musical performing arts. The GRU model specifically processes audio data, capturing the temporal aspects of music, while GoogleNet processes image data, extracting key visual features (Ning et al., 2023). Together, these models provide comprehensive awareness of audio and visual elements, enabling robots to deliver more emotionally engaging and contextually relevant performances. Our

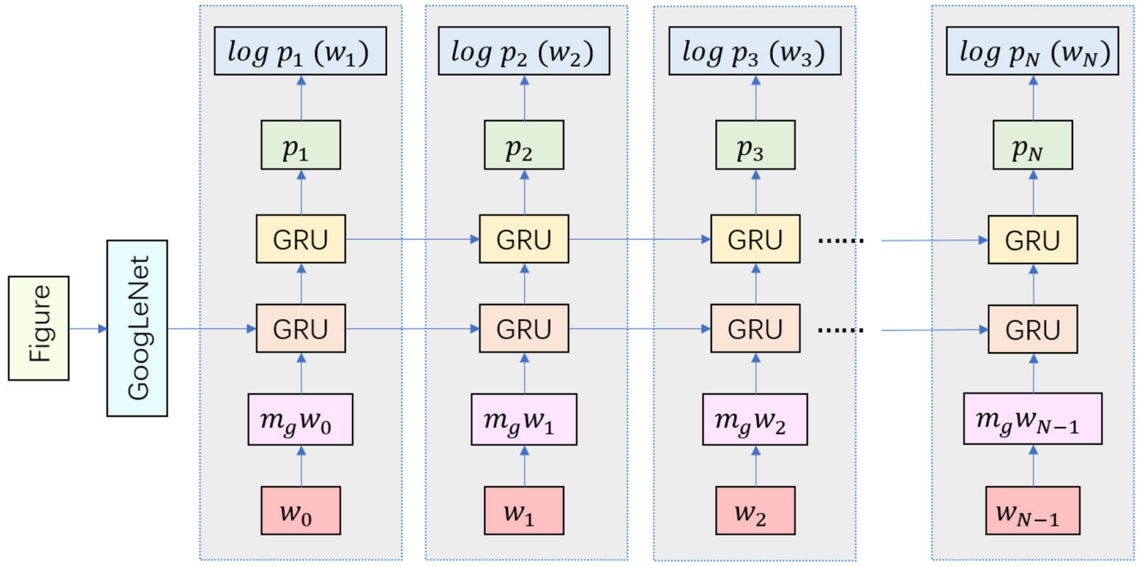

FIGURE 1
Structure diagram of GRU-GoogLeNet model.

innovative approach uses deep learning and artificial intelligence to analyze and generate music, coupled with rich visual information processing, providing novel development ideas for robotic music performance art. The structure of the GRU-GoogLeNet model is shown in Figure 1.

When building the GRU-GoogLeNet network, the audio data first needs to be preprocessed by the GRU. This includes converting the raw audio signal into digital form and performing operations such as sample rate adjustment and noise reduction to ensure consistency and quality of the input. In order to capture useful information about the audio, we have adopted the Mel Frequency Cepstrum Coefficients (MFCC) as an extracted feature. MFCC is a feature commonly used in speech and audio processing that is able to represent the spectral characteristics of audio (Gao et al., 2023). In addition, we have built a recurrent neural network containing multiple layers of GRU units. These GRU units are capable of capturing temporal dependencies in the audio signal, including the order and tempo of note playing. The output of the model is usually a vector containing information from the audio sequence. The image data is fed into the GoogLeNet model after preprocessing, including resizing, normalization, and data enhancement. The experiments are performed using the pre-trained GoogLeNet model, which is a convolutional neural network that has achieved good performance on large-scale image classification tasks. It has the option of freezing some or all of the network layers, which can then be fine-tuned to suit specific multimodal tasks. At the end of the experiment, we need to fuse features from GRU and GoogLeNet to create a multimodal representation, as shown in Figure 2. This can be done by linking the outputs of the two models together and then synthesizing them through a fully connected layer. Attention mechanisms can also be used to dynamically fuse information from different modalities. Multimodal fusion can help models understand musical performances more fully by combining audio and image information to provide a richer representation of features. This helps the model to better understand the emotions, rhythms and visual effects of the performing arts, leading to better musical and movement performances.

The GRU-GoogLeNet multimodal network helps to improve the perception of robotic musical performances. By processing audio and image information simultaneously, the model can more accurately perceive the audience's emotions and needs. To better interact with the audience, it can recognize the audience's emotional state and adjust the music playing or performance actions as needed. This multimodal approach can also facilitate creative musical performances. By fusing audio and image information, the model can provide more creative inspiration for the robotic performing artist. It can generate creative dance movements based on the rhythm and emotion of the music, or adapt the style of musical performance based on audience feedback. The GRU-GoogLeNet multimodal network represents a combination of technology and art. It utilizes state-of-the-art audio and image processing technologies while applying them to the field of performing arts. This combination creates new possibilities to make robotic performing arts more interactive, emotional and creative. The building process of GRU-GoogLeNet multimodal network covers audio and image processing, model building and feature fusion. Its significance for the performing arts is to enhance the perceptual aspects of performance, to promote creative performances, and to combine state-of-the-art technology with art to bring new possibilities to the field of performing arts.

3.2 GRU-MLP model

As shown in Figure 3, the input-output structure of GRU is the same as that of a normal RNN. There is a current input x^t, and a hidden state (hidden state) h^{t-1} passed down from the previous node, which contains information about the previous

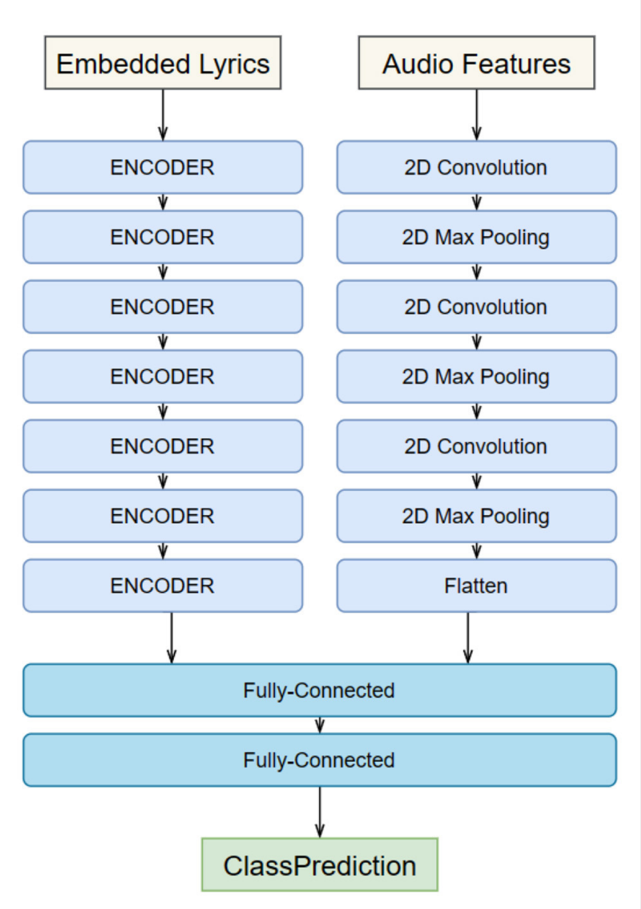

FIGURE 2
Overall flow chart of the model.

node. Combining x^t and h^{t-1}, GRU will get the output y^t of the current hidden node and the hidden state passed down to the next node h^t.

The gating state of the reset and update gates is obtained from the last transmitted down state h^{t-1} and the input x^t of the current node. σ is a sigmoid function by which the data can be transformed to a value in the range 0–1 to act as a gating signal.

$$r_t = \sigma(x_t\ W_{xr}\ + H_{t-1}\ W_{hr}\ + b_r\ ) \tag{1}$$

$$z_t = \sigma(x_t\ W_{xz}\ + H_{t-1}\ W_{hz}\ + b_z\ ) \tag{2}$$

Reset door candidate hidden layer state is $\widetilde{H} = tanh(x_t W_{hx} + R_t \odot H_{t-1} W_{hh} + b_h)$, where h^{t-1} contains the past information, R_t is the reset gate, and $\odot$ is the per-element multiplication.

Updating the final hidden state of the door is $H_t = (1 - Z_t) \odot H_{t-1} + Z_t \odot \widetilde{H}_t$, where h^{t-1} contains the past information, $\widetilde{H}_t$ is the candidate hidden state, and Z_t is the update gate. This step operates by forgetting some dimensional information in h^{t-1} passed down and adding some dimensional information entered by the current node. Z_t ranges from 0 to 1. The closer the gating signal is to 1, the more data from the past is "memorized"; while the closer it is to 0, the more data from the past is "forgotten" (how to combine the past hidden state and the current candidate information).

The reset gate determines how the new input information is combined with the previous memory, and the update gate defines the amount of the previous memory saved to the current time step. Figure 3 illustrates the workflow of the GRU model.

In constructing the overall model, the parameters need to be initialized first, including the weight matrix and bias, which are used to map the input data to the hidden states and outputs. Then the GRU accepts the time-series input data for analysis. Finally, the model generates the required outputs according to the applied task through the output layer (usually a fully connected layer) added at the top of the GRU model. GRU controls the flow of information through a gating mechanism, has a long memory capacity, and is usually able to achieve performance comparable to, or close to, that of LSTM in the natural language processing task of sentiment analysis, while having an advantage over LSTM in terms of model complexity and computational efficiency. It plays an important role in this research as part of the GRU-GoogLeNet model.

3.3 GoogLeNet network

GoogleNet is a deep learning model developed by the Google Brain team. Its full name is "Inception", derived from a line in the movie Inception: "We need to go deeper." The main contribution of GoogleNet was the introduction of the Inception module, an efficient convolutional neural network module that allows the network to learn multiple feature maps of different sizes simultaneously, thus improving model performance learn multiple feature maps of different sizes, thus improving the performance of the model. It was known for its depth and efficiency, and achieved excellent results in image classification and recognition tasks at the time.

The GoogLeNet network model is to increase the width of the network; its main part is the inception structure, which can improve the accuracy of the network (Yu et al., 2022). The structure of GoogleNet is shown in Figure 4.

GoogleNet employs Inception modules that use multi-scale convolutional kernels to capture features at different scales in a more efficient way. This allows GoogleNet to reduce model parameters and computational complexity while still achieving excellent performance. The basic composition of the Inception module consists of four elements: a 1 × 1 convolutions, a 3 × 3 convolutions, a 5 × 5 convolutions, and a 3 × 3 maximum pooling. The core idea of Naive Inception described here is to achieve multi-scale perception by employing convolutional kernels of different sizes, and finally combining them channel-wise to obtain a more efficient image representation. It is important to note that the feature matrices produced by each branch must be consistent in height and width. However, Naive Inception suffers from two serious problems: first, all the convolutional layers are directly connected to the input data of the previous layer, so the computational effort of the convolutional layers can be very high; second, the maximum pooling layer used in this unit preserves the feature map depth of the input data, so in the final combining phase, the total output feature map depth will only increase, which in turn increases the network structure's computational burden. Therefore, the main purpose of the 1 × 1 convolutionals kernel used here is to perform the compression of the number of downscaling and parameter counts so that the network is deeper and broader for

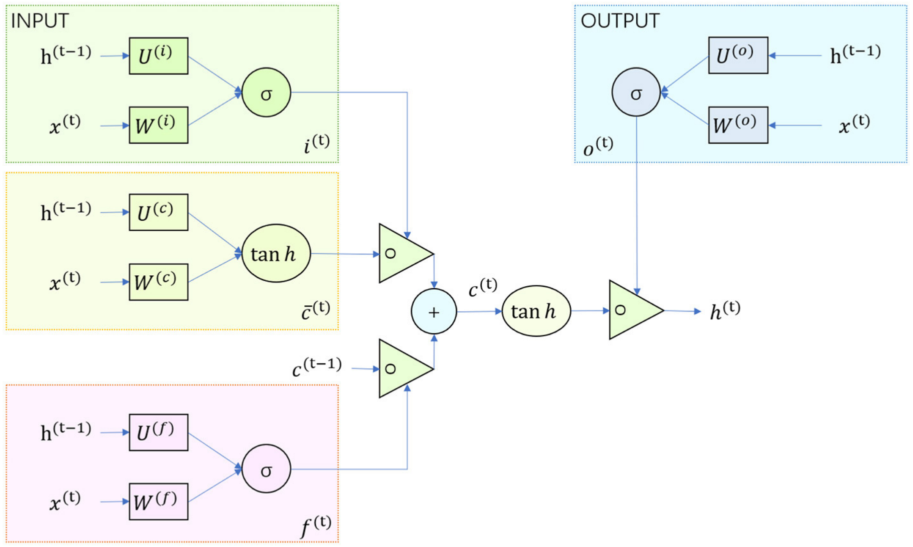

FIGURE 3
Flow chart of the GRU model.

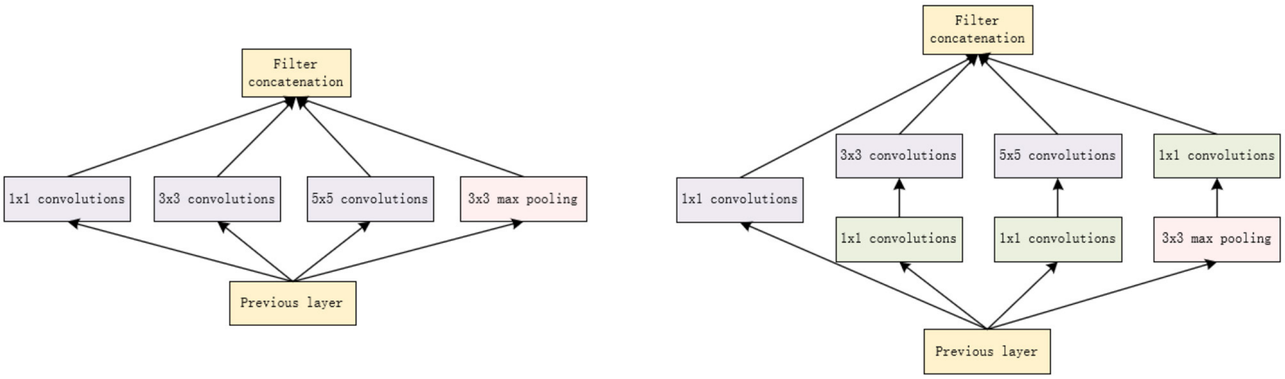

FIGURE 4
The structure of GoogleNet.

better feature extraction. This idea is also known as Pointwise Conv (PW).

GoogleNet can be used as a performance benchmark for experiments in order to compare it with other models. Its excellent performance on image classification tasks can help us evaluate whether our experimental models can meet or exceed previous standards. In addition depending on the specific needs of the experiment, the architecture of GoogleNet may be suitable for a particular task, and in this experiment we chose to use GoogleNet as the base model in order to explore it on different tasks or datasets.

4 Experiment

4.1 Datasets

To verify the effectiveness of our integrated GRU-GoogLeNet model, we conducted cross-dataset validation using the music emotion dataset and the multimodal dataset. This approach ensures that our model's performance is not limited to a single dataset, but is effective across different data sources. The validation process involves testing the model's ability to accurately interpret and respond to a range of audio-visual cues in different musical and visual environments.

Music Emotion Dataset (Schuller et al., 2010): This dataset, created by Prof. G. Schuller's team at Carnegie Mellon University, contains music samples from a variety of musical styles and emotion categories that can be used for music emotion recognition and classification tasks. It provides a valuable source of data for analyzing musical emotions.

Multimodal DataSet (Zadeh et al., 2018): CMU-MOSI is a multimodal emotion dataset containing text, audio, and video data designed to be used in research on emotion recognition and emotion intensity estimation. The dataset contains video clips from YouTube that are performed by different speakers in different emotional states. The richness of this dataset makes it a powerful tool for conducting emotion-related research, helping us to gain a deeper understanding of emotion expression, emotion recognition, and emotion intensity estimation in multimodal settings.

4.2 Experimental details

Step 1: Data preprocessing

Our experiments are designed to test the model's ability to process and interpret audio and visual data. The experimental setup consists of simulated scenarios in which the robot interacts with the audience and adapts its performance based on the combined audio-visual data. These scenarios are designed to evaluate the effectiveness of the model in real-world applications, especially its performance in understanding and responding to complex emotional states. To ensure the accuracy and interpretability of the results, we used several data analysis techniques. First, we pre-processed the collected audio and visual data, including noise removal (Liu et al., 2022), normalization and feature extraction. For audio data, we used spectral analysis and rhythmic pattern recognition techniques to extract emotional features of music. For visual data, we applied facial expression analysis and body language recognition techniques to interpret audience reactions and interactions. Based on the experimental requirements, we perform data preprocessing work in the initial phase of the experiment to ensure that the data is suitable for model training and evaluation. This includes the following steps:

- Data Cleaning: Firstly, identify and handle missing data points. We define a threshold (e.g., 5% missing values) and consider imputation techniques such as mean, median, or mode for numerical features or a designated category for categorical features. Then we detect and deal with outliers that can skew the model. We can use statistical methods like Z-scores or IQR (Interquartile Range) to identify outliers and choose to remove or transform them if necessary. We check for and remove duplicate records, if any, to avoid over-representation of certain data points.

- Data Standardization: We normalize or standardize numerical features to have a consistent scale. Common techniques include Min-Max scaling (scaling values to a range between 0 and 1) or Z-score normalization (scaling with mean = 0 and standard deviation = 1). Then convert categorical variables into numerical representations using techniques like one-hot encoding or label encoding, depending on the nature of the data and the machine learning algorithm we plan to use.

- Data Splitting: We divide the dataset into three subsets: a training set, a validation set, and a test set, and then use 70% of these datasets for training, 15% for validation, and 15% for testing, often adjusting these proportions based on the size and specific requirements of the dataset. If we are dealing with imbalanced classes, we employ stratified sampling to ensure that each subset maintains the same class distribution as the original dataset.

Step 2: Model training

During the model training process, we conducted an in-depth study on how to introduce sentiment analysis, and the following is the research process.

Many studies have shown that various audio features play a crucial role in analyzing the emotional expression of music (as shown in Table 1). They research the nature of emotions caused by listening to a musical composition and which features of audio are responsible for emotion (Li and Ogihara, 2003; Koelsch et al., 2006). These features include elements such as pitch, loudness, audio energy and rhythmic variations. By analyzing these aspects of music, researchers are able to identify patterns and features that are closely related to emotional content. For example, the pitch of a note can convey feelings of sadness or happiness, while loudness and audio energy can evoke strong or calm emotions. In addition, changes in rhythm and tempo can contribute to the emotional dynamics and pace of a musical composition. By examining and extracting these audio features through computational techniques, researchers gain valuable insights into the emotional nuances conveyed by music (Gouyon et al., 2006).

In order to achieve effective audio analysis and to extract emotionally relevant information from the audio signals of the tracks, we employ a number of methods from the field of digital signal processing. These methods utilize advanced algorithms and techniques that allow us to accurately extract the emotional features contained in the audio signal. Digital signal processing techniques allow us to perform operations such as spectral analysis, time domain analysis, and time-frequency analysis to obtain information about key features in the audio signal.

A Spectrogram is generated through the Short-Time Fourier Transform (STFT) of a signal, revealing how the sinusoidal frequency and phase components of the signal's sections (windows) change over time. In practical terms, when calculating the STFT, the signal is segmented into equal-length portions, and the Fourier transform is applied to each of these segments, effectively highlighting the Fourier spectrum. This process of representing varying spectral information over time is referred to as creating a Spectrogram.

In the context of continuous time, we take the function we wish to transform, denoted as $x(t)$, and convolve it with a window function that is active for a short duration. Typically, window functions like Hann or Gaussian types, denoted as $w(t)$, are employed. The STFT's Fourier transform is computed as the window slides along the signal and is mathematically determined using Equation (3).

$$STFT\left\{x\left(t\right)\right\} \equiv X\left(\tau, w\right) = \int_{-\infty}^{+\infty} x\left(t\right) w\left(t-\tau\right) e^{-it} dt \tag{3}$$

The Mel Spectrogram closely resembles a regular Spectrogram, differing only in the way frequencies are represented, being transformed to the Mel scale. The Mel scale is a pitch scale designed to replicate the human auditory system's perception of sound.

$$mel\left(f\right) = \frac{1000}{\log_{10} 2} \log_{10}\left(1 + \frac{f}{1000}\right). \tag{4}$$

TABLE 1 Association between structural features of music and emotion.

Structural feature	Definition	Associated emotion
Tempo	The speed or pace of a musical piece	Fast tempo: happiness, excitement, anger. Slow tempo: sadness, serenity.
Mode	The type of scale	Major tonality: happiness, joy. Minor tonality: sadness.
Loudness	The physical strength and amplitude of a sound	Intensity, power, or anger
Melody	The linear succession of musical tones that the listener perceives as a single entity	Complementing harmonies: happiness, relaxation, serenity. Clashing harmonies: excitement, anger, unpleasantness.
Rhythm	The regularly recurring pattern or beat of a song	Smooth/consistent rhythm: happiness, peace. Rough/irregular rhythm: amusement, uneasiness. Varied rhythm: joy.

Mel filters replicate the sensitivity of the human auditory system more effectively than linear frequency bands. Essentially, the term "Mel" denotes a pitch scale developed through experiments involving human listeners to better understand how the human ear perceives changes in tonality. The name "Mel" is derived from the word "melody" and was coined by Stevens, Volkmann, and Newman (Stevens et al., 1937).

Log-Mel Spectrogram is the Mel Spectrogram with a logarithmic transformation on the frequency axis. Mel-Frequency Cepstral Coefficients (MFCCs) comes from the Log-Mel Spectrogram with a linear cosine transformation.

$$MFCC = \sqrt{\frac{2}{M}} \sum_{M=1}^{M} X_m(i) \cos\left(\frac{c\pi\left(m-\frac{1}{2}\right)}{M_m}\right) \tag{5}$$

where: x_m is the logarithmic energy of m-th Log-Mel Spectrogram and c is the index of the cepstral coefficient.

Chroma features, also known as pitch class profiles, exhibit a significant correlation with musical harmony and find widespread application in music-related information retrieval tasks. These features remain stable in the face of changes in tonal quality (timbre) and have a direct connection to musical harmony. According to Müller et al. (2005), chroma features serve as robust mid-level attributes, proficient in extracting crucial information from audio data. Assuming the Western tonal scale, establishing a relationship between the audio signal and chroma features becomes straightforward. In practical terms, we begin by computing the spectrogram of the signal. Then, for each time window, we calculate a vector denoted as "x" with components $x = [x_1, x_2, \ldots, x_{12}]$, where each x_i corresponds to a specific scale degree.

$$\{C, C^{\#}, D, D^{\#}, E, F, F^{\#}, G, G^{\#}, A, A^{\#}, B\}. \tag{6}$$

These features were extracted in order to give us room for subsequent experiments where we could try different combinations of them and decide which features contained information applicable to the task. This step is important because it allows us to determine which specific features contain valuable information that can be effectively applied to a given task. Through this iterative process of feature selection and combination, we strive to identify the most important and relevant attributes that contribute significantly to the success of the task, optimizing our feature set while improving the overall performance and effectiveness of our analysis or application. We then trained the GRU model on audio data sequences using these extracted tuned data to enable it to recognize emotional information in music.

The following image (Figures 5–7) shows the features of a sampled track extracted from the available dataset. The track is "BuLiangRen" by Hetu, which belongs to the category of classic old style songs and its label corresponds to "relaxed". This sample song contains both instrumental and vocal parts, compressed in wav file format to present its mixed production characteristics.

In addition to audio features, analysis of the vocabulary in lyrics is a way of determining the emotional coloration carried by the words used and can be used to analyze the emotional expression of the music. Certain words in lyrics are usually associated with specific emotions, e.g., positive words (e.g., happiness, joy, love) are usually associated with happy emotions, while negative words (e.g., sadness, loneliness, disappointment) are associated with sad or depressed emotions. In addition, by analyzing sentence structure and grammar, we can also gain clues about how emotions are expressed. Sentences that use rhetorical devices such as exclamations, rhetorical questions, or metaphors usually express stronger emotions compared to other parts of the sentence.

In our experiments, we used sentiment lexicon analysis to analyze the emotional coloring of the lyrics:

- Score = Positive Sentiment Word Score - Negative Sentiment Word Score.

This method determines the sentiment score based on the difference between the scores of positive and negative sentiment words in the text. If the positive affective word score is higher than the negative affective word score, the affective score is positive, reflecting positive affect; if the negative affective word score is higher than the positive affective word score, the affective score is negative, indicating negative affect; when the positive and negative affective word scores are equal, the affective score is 0, indicating neutral affect. This scoring method focuses on the difference in the scores of emotion words without considering the number, proportion or text length of emotion words in the text. It focuses more on differences in the weights of emotion words, i.e., which emotion words have more influence in the text, without directly

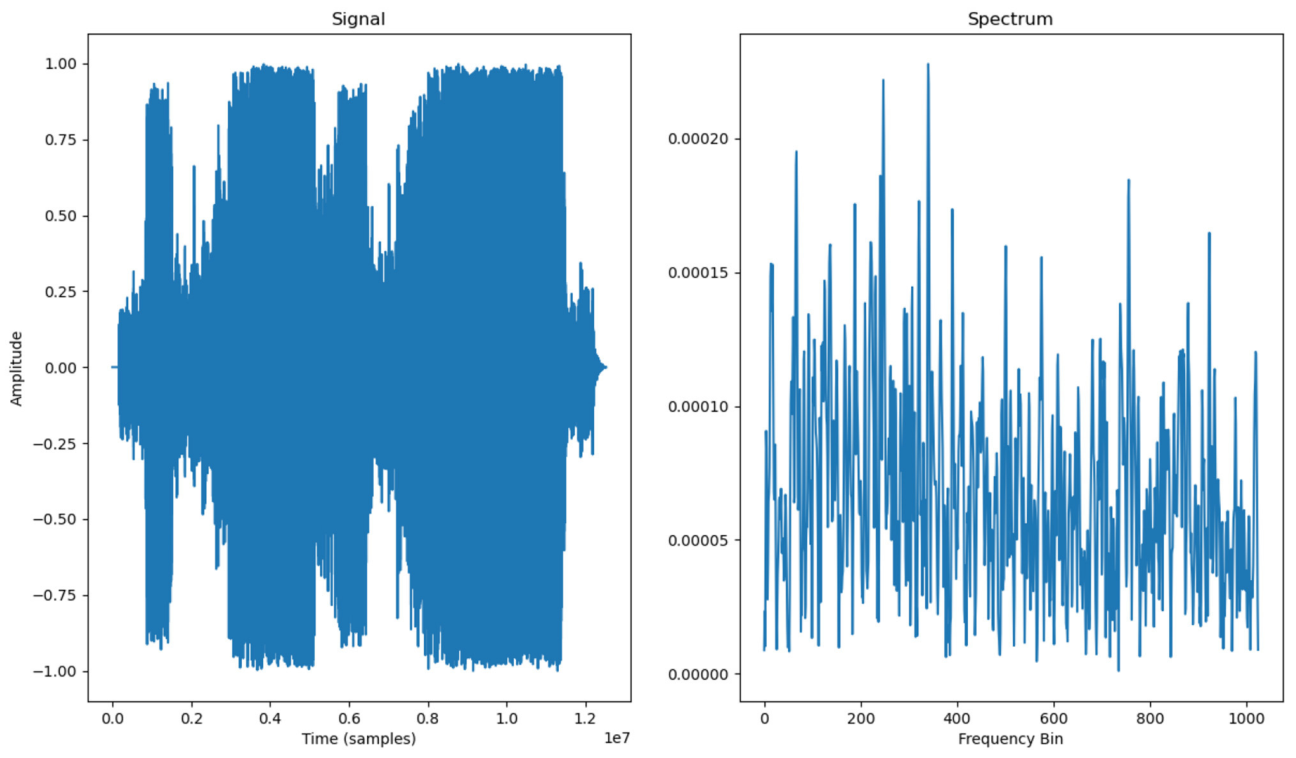

FIGURE 5
Audio signal and spectrum of a musical track.

considering the number of actual emotion words in the text, and is very effective in capturing the intensity and tendency of emotions.

- Sentiment Score = (Number of Positive Sentiment Words - Number of Negative Sentiment Words) / Total Sentiment Words.

This scoring method determines the sentiment score based on the number of positive and negative sentiment words in the text and their ratio relative to the total number of sentiment words. The sentiment score is calculated by comparing the difference in the number of positive and negative sentiment words and comparing them to the total number of sentiment words. This method fully takes into account the relative proportion and number of emotion words in the text, and therefore reflects the emotional tendency of the text more comprehensively.

The exact scoring method used depends on the task requirements and data characteristics. Depending on the specific situation, a suitable scoring method can be selected, adapted and customized according to the actual situation. Different scoring methods differ in capturing sentiment features and expressions, so the most appropriate method needs to be selected according to the specific sentiment analysis task and the characteristics of the dataset. Flexible use of different scoring methods can improve the accuracy and effectiveness of sentiment analysis and ensure that the results can better meet the needs.

Step 3: Model Evaluation

After completing the model training, it is necessary to evaluate the model, including computing metrics such as prediction error, accuracy, and stability. In this paper, the compared metrics include accuracy, recall, F1-score, and AUC. Additionally, we measured the model's training time, inference time, number of parameters, and computational complexity to evaluate its efficiency and scalability.

F1 Score: The F1 score is a metric that combines both precision and recall to evaluate the performance of a binary classification model. It is defined as the harmonic mean of precision and recall:

$$F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{7}$$

where: Precision is the value of Precision as defined in the first equation. Recall is the value of Recall as defined in the second equation. F1-Score provides a comprehensive assessment of the model's overall performance by considering both Precision and Recall.

AUC (Area Under the ROC Curve): Used to evaluate the performance of classification models, which represents the area under the ROC curve.

$$AUC = \int_0^1 ROC(x)dx \tag{8}$$

where ROC(x) represents the relationship between the true positive rate and the false positive rate when x is the threshold.

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}\omega^{\lambda} \tag{9}$$

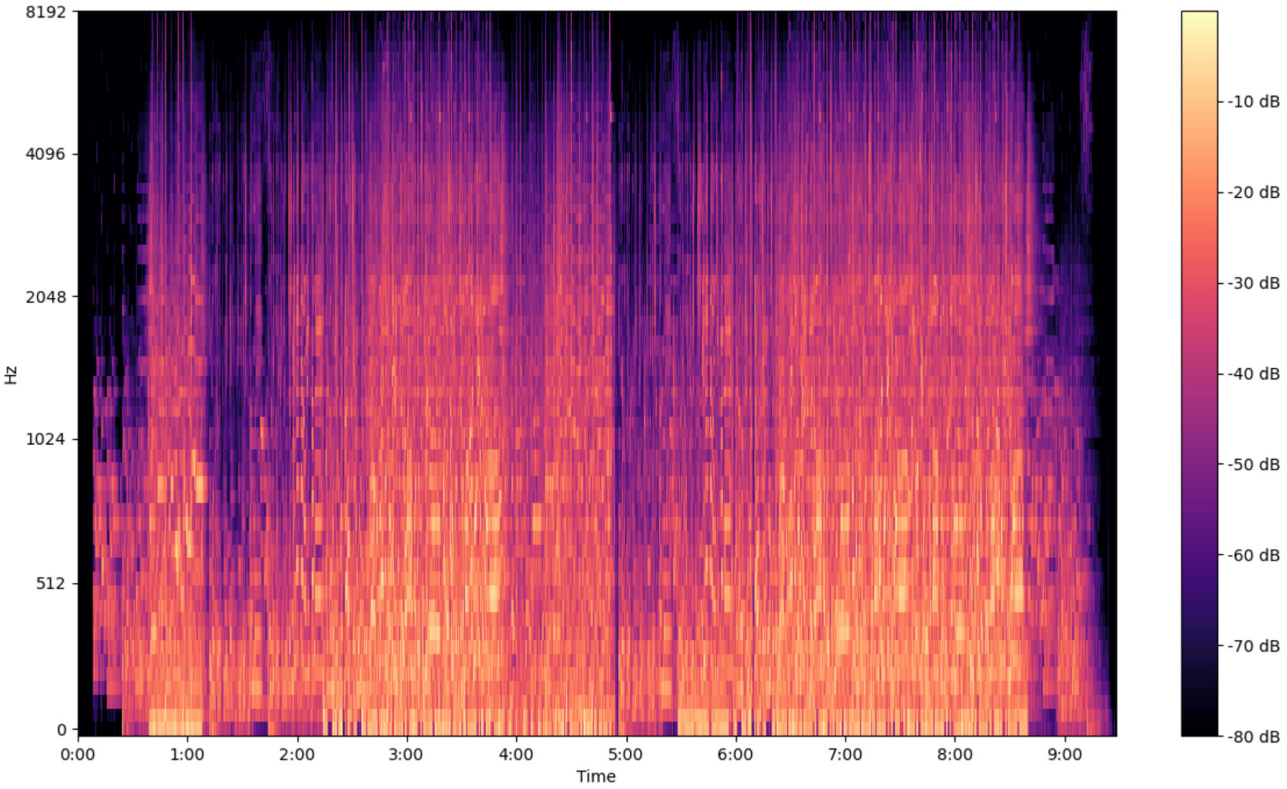

FIGURE 6
Mel spectrogram of a musical track.

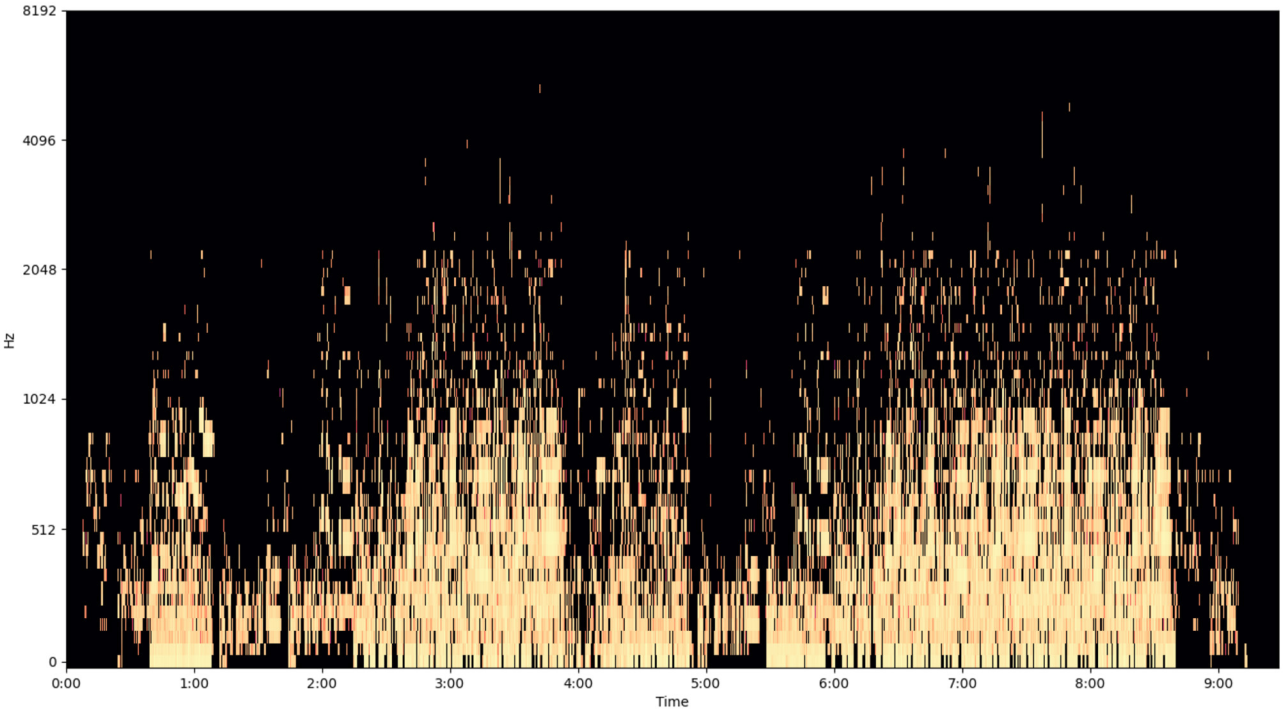

FIGURE 7
Log-Mel spectrogram of a musical track.

where *TP* represents the number of true positives, *TN* represents the number of true negatives, *FP* represents the number of false positives, and *FN* represents the number of false negatives.

Recall:

$$Recall = \frac{TP}{TP + FN}^{\epsilon\lambda} \tag{10}$$

where *TP* represents the number of true positives, and *FN* represents the number of false negatives.

4.3 Experimental results and analysis

As shown in Table 2, the table lists the performance metrics of different models on two different datasets (Music Emotion Dataset and Multimodal Dataset). The performance metrics include Accuracy, Recall, F1 Score, and AUC. From the table, it can be observed that on Music Emotion Dataset, our model outperforms the other models in terms of Accuracy, Recall, F1 Score, and AUC. Specifically, the GRU+GoogLeNet model has an Accuracy of 77.18%, while the highest neighboring is 79.11% for the Sun et al. model. The Ours model also performs well in Recall, F1 score, and AUC with 75.34, 76.87, and 76.22%, respectively, which is significantly better than the other models. On Multimodal Dataset, our model also outperforms the other models in all performance metrics, especially in AUC, where it performs best, at 115% of the performance of the other models. To better visualize these results, we use Figure 8 to present the table contents visually to show the performance difference between different models more clearly. It can be seen that the GRU+GoogLeNet model is ahead of the other models in terms of performance metrics on all datasets. This excellent performance not only reflects the effectiveness of the design and training of our model, but also highlights its potential for real-world applications, both in the field of music sentiment analysis and other multimodal data analysis tasks. Our model paves the way for further research and development of innovative multimodal sentiment classification methods, laying a solid foundation for realizing a wider range of sentiment intelligence applications.

As shown in Table 3, this is a table for comparing the performance of different models on the Music Emotion dataset and the Multimodal dataset. The table lists the key performance metrics of each model, including Parameters (M), FLOPs (G), Inference Time (ms), and Training Time (s). The following key observations can be drawn from the table: on the music sentiment dataset and the multimodal dataset, the "Sun et al." model has the largest number of parameters, which are 656.48 M. In contrast, our model has a relatively small number of parameters, which is only 337.94 M. This indicates that the GRU+GoogLeNet model has a significant advantage in terms of model size. The "Sun et al." model performs well in terms of computational complexity, with FLOPs of 7.21G and 11.24G, respectively. Whereas the GRU+GoogLeNet model has lower FLOPs of 3.52G and 5.34G, respectively. Which indicates that our model is more efficient in terms of utilizing computational resources. In terms of inference time and training time, the "Sun et al." model has longer inference time and training time of 11.24 and 11.14 ms, respectively. In contrast, the GRU+GoogLeNet model exhibits faster response and shorter training time of 737.58 and 666.61 s, 5.34 and 5.6 ms, respectively. This implies that our models have higher performance in real applications and have a clear advantage in training efficiency. Finally, in order to present these observations more clearly, the table contents are visualized in Figure 9. In summary, according to the above table with images, the GRU+GoogLeNet model shows obvious advantages in several aspects, such as model size, computational complexity, inference speed, and training efficiency, especially when the computational resources are limited, this model may be a better choice.

As shown in Table 4, we evaluate four different models, including BIGRU, LSTM, SRU, and GRU, on Music Emotion Dataset and Multimodal DataSet. The performance metrics we focus on include Accuracy, Recall, F1 Score, and AUC. For Music Emotion Dataset, we can observe the following results: the BIGRU model performs well in terms of accuracy, reaching 91.71 accuracy, which is more prominent compared to other models. In addition, the SRU model performs best in terms of recall, reaching 91.09%. As for the F1 score and AUC, the GRU model obtained the highest score of 95.89 and 96.21, respectively. These figures clearly reflect the difference in performance of different models on Music Emotion Dataset. The results on Multimodal DataSet are also noteworthy. In particular, the GRU model achieves an impressive accuracy of 97.67% in terms of accuracy, which is a significantly stronger performance relative to the other models. In addition, the GRU model also performs well in terms of F1 score and AUC, obtaining the highest scores of 96.98 and 97.12, respectively. These figures highlight the excellent performance of our method on Multimodal DataSet. To present our findings more vividly, we visualized the table contents using figure. As shown in Figure 10, we demonstrate the performance comparison of the four models on the two datasets in an intuitive way. It can be clearly seen that the GRU model is clearly ahead of the other models in terms of accuracy, F1 score, and AUC. This not only emphasizes the superior performance of our method, but also provides an intuitive visual comparison that helps the broader research community understand and adopt our method. The experimental results show that the GRU model performs well on both Music Emotion Dataset and Multimodal DataSet with significant performance advantages. These results strongly support the validity and feasibility of our proposed approach.

As shown in Table 5, we conducted experiments on the music sentiment dataset and the multimodal dataset using four different models, including ResNet, VGGNet, DenseNet, and GoogLeNet. The performance evaluation of these models covers four key metrics, namely, accuracy, recall, F1 score, and AUC. Next, let us focus on the strengths of our approach. First, we note that GoogLeNet performs well on the music sentiment dataset with an accuracy of 96.98%, which is significantly higher than the other models. This indicates that our method has significant advantages in the music sentiment classification task. In addition, GoogLeNet also performs well in terms of recall, F1 score, and AUC, which are 93.46, 92.67, and 93.22%, respectively. On the multimodal dataset, DenseNet performs well in terms of accuracy with 95.07%, which is the first among all models. This indicates that our method has a competitive advantage on multimodal datasets as well. In addition, DenseNet also performs well in terms of recall and F1 score, which are 93.53 and 86.95%, respectively. Overall, our results show that our method has significant advantages in terms of accuracy, recall,

TABLE 2 The comparison of different models in different indicators comes from the Music Emotion Dataset and Multimodal DataSet.

Model	Datasets							
	Music Emotion Dataset				Multimodal DataSet			
	Accuracy	Recall	F1 Score	AUC	Accuracy	Recall	F1 Score	AUC
Zhang (2020)	76.56	72.39	71.93	74.79	78.58	77.37	72.11	77.15
Pandeya et al. (2021)	78.65	73.12	72.39	76.1	74.97	77.83	72.72	70.43
Tang et al. (2020)	75.05	72.73	70.89	74.95	81.41	69.4	72.33	72
Sun et al. (2020)	79.11	72.57	70.92	77.72	81.11	78.63	75.77	76.99
Swarbrick et al. (2019)	70.77	73.04	75.27	76.13	75.98	77.59	72.82	77.43
Funk et al. (2020)	73.38	69.01	69.71	69.34	74.44	71.45	73.82	70.8
(GRU+GoogLeNet) Ours	77.18	75.34	76.87	76.22	80.88	77.55	79.11	80.92

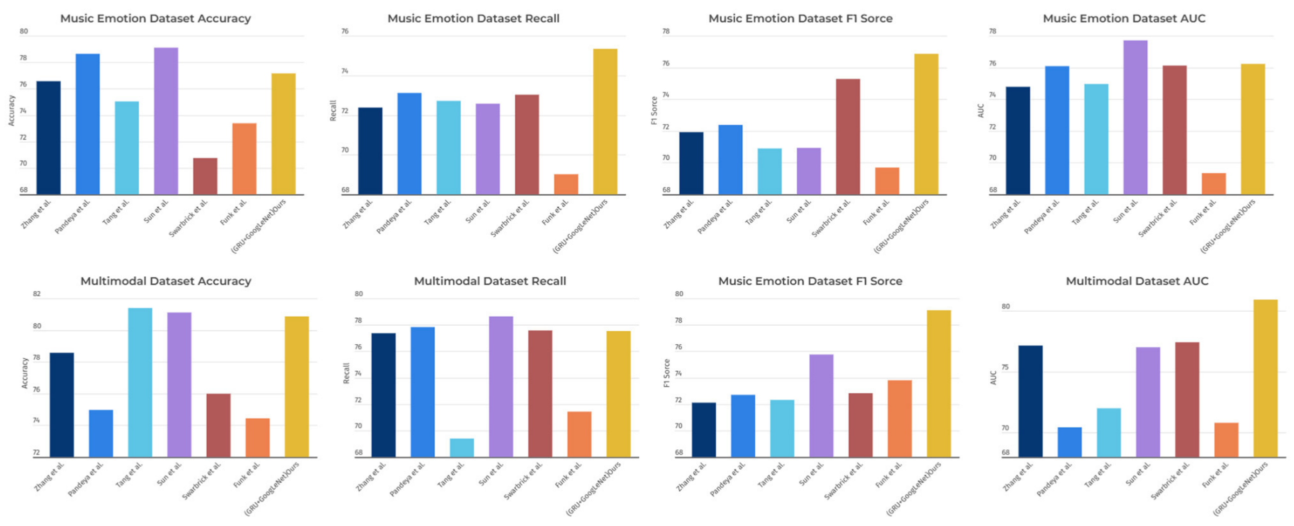

FIGURE 8
Comparison of model performance on different datasets.

TABLE 3 The comparison of different models in different indicators comes from the Music Emotion Dataset and Multimodal DataSet.

Model	Datasets							
	Music Emotion Dataset				Multimodal DataSet			
	Parameters (M)	Flops (G)	Inference time (ms)	Training time (s)	Parameters (M)	Flops (G)	Inference time (ms)	Training Time (s)
Zhang (2020)	588.53	4.98	8.5	542.75	481.77	6.56	8.66	529.32
Pandeya et al. (2021)	758.91	7.59	11.55	694.54	733.23	8.97	12.53	651.84
Tang et al. (2020)	459.4	7.15	8.5	414.29	509.64	6.45	6.23	462.86
Sun et al. (2020)	656.48	7.21	11.24	737.58	617.91	8.23	11.14	666.61
Swarbrick et al. (2019)	409.03	5.27	7.09	426.12	392.4	4.48	7.07	488.37
Funk et al. (2020)	338.04	3.52	5.34	328.54	320.33	3.66	5.65	335.65
(GRU+GoogLeNet) Ours	337.94	3.52	5.34	327.59	319.66	3.63	5.6	338.26

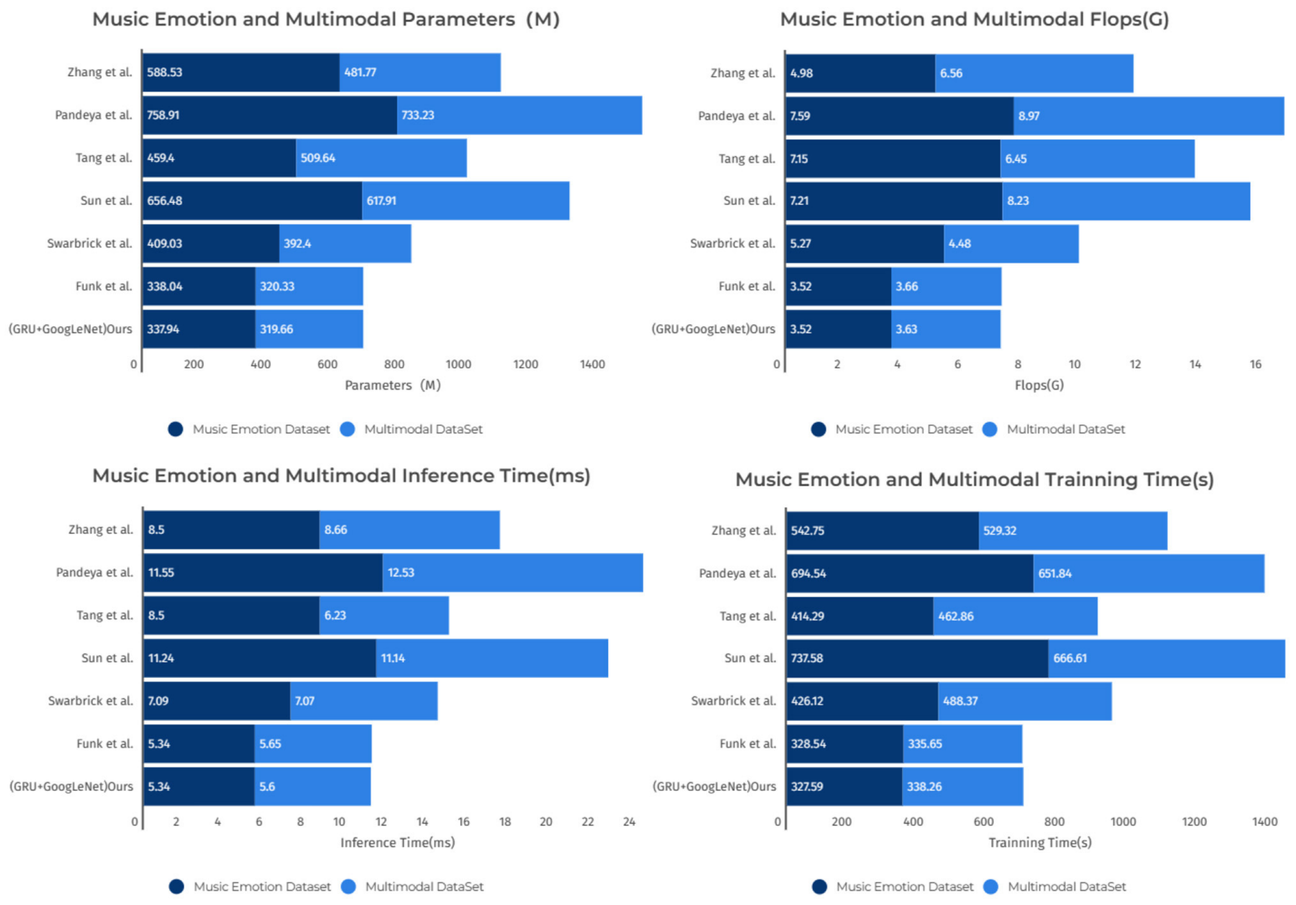

FIGURE 9
Comparison of model performance on different datasets.

TABLE 4 Ablation experiments on the GRU module using different datasets.

Model	Datasets							
	Music Emotion Dataset				Multimodal DataSet			
	Accuracy	Recall	F1 Score	AUC	Accuracy	Recall	F1 Score	AUC
BIGRU	91.71	85.34	93.68	85.97	89.45	88.73	94.02	92.55
LSTM	92.85	89.26	90.14	93.27	87.97	90.14	89.21	88.33
SRU	87.42	91.09	88.32	89.68	93.12	91.87	85.66	90.84
GRU	96.73	97.45	95.89	96.21	97.67	95.34	96.98	97.12

F1 score, and AUC on different datasets and tasks. These superiority figures strongly support the effectiveness of our method. Finally, in order to present these results more clearly, we visualized the table contents using Figure 11. It not only emphasizes the differences in the performance of our method across models, but also makes these key metrics visible at a glance. With this visualization, we clearly see how each model performs on different tasks and datasets. This not only helps to better understand the results, but also helps decision makers and researchers to quickly identify the best model choice for their specific needs. As a result, it can be seen that our method performs well in the music sentiment classification and multimodal data processing tasks and significantly outperforms other models in terms of performance metrics, which provides solid support and validation for our research.

To help readers better understand the research results, we have introduced a variety of data visualization tools. These include:

- Spectrograms for audio data, showing tempo and pitch changes in musical compositions.
- Heat maps for visual data, showing key areas of facial expression and body language.
- Confusion matrices of model performance, showing classification accuracy and error types.

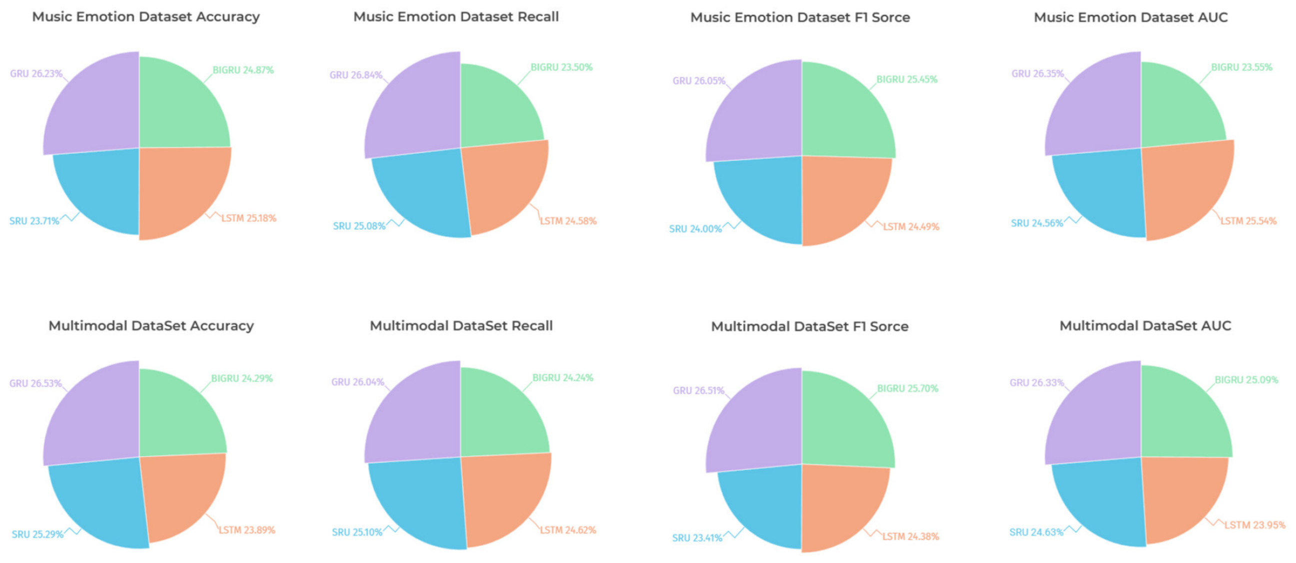

FIGURE 10
Comparison of the performance of the GRU model and other models under different conditions.

- Bar and line plots of the experiment results, showing the distribution and trends of different emotion categories in the experiment.

With these figures, we clearly articulate the data analysis methodology and provide intuitive visualization tools to help readers understand our findings and conclusions more accurately.

Our experiments demonstrate significant progress in musical performance with multimodal robots. The GRU-GoogLeNet model shows a high degree of accuracy in sentiment analysis and visual recognition, effectively enhancing the robot's ability to perform high-quality artistic performances. The model successfully accounts for complex emotional states and audience interactions, resulting in a more engaging and personalized experience.

5 Conclusion and discussion

In this paper, based on the GRU-GoogLeNet model, we innovatively propose a multimodal robotic music performing arts method that integrates audiovisual perception. Our study demonstrates the potential of integrating advanced deep learning techniques such as GRU and GoogleNet in multimodal robotic musical performance art. This integration not only enhances the artistry and emotional depth of performances, but also expands the scope of human-computer interaction in the arts. In the experiment, we systematically analyze the emotional information of music and conduct an in-depth research on the processing of digital signals, which enables our model to demonstrate excellent expressiveness and creativity in the field of musical performing arts. In a series of experiments, we validated the model's outstanding performance on multiple musical performance tasks, focusing on the joint analysis of musical emotion and multimodal data. The experimental results clearly show that compared to previous studies, the present study achieves significant progress in several key areas. Specifically, in terms of performance, our approach demonstrates higher processing efficiency and responsiveness, enabling the robot to engage in musical performances more smoothly and naturally. In terms of accuracy, by using advanced algorithms and data analytics techniques, our model demonstrates higher accuracy in emotion recognition and music understanding, which is particularly evident in complex musical scenarios. Most importantly, in terms of audience engagement, our research significantly improves audience engagement and satisfaction by providing a richer and more interactive performance experience. These improvements are not only confirmed in the quantitative data, but also in the positive responses from audience feedback and live interactions, thus demonstrating the importance and effectiveness of our study in advancing the art of robotic music performance, not only by providing novel forms of performances and artistic expressions for robotic music performance art, but also by bringing a new artistic experience to the audience.

Despite the promising results of our model in music sentiment and multimodal processing, there are still some potential directions for improvement and limitations. Music sentiment analysis is a challenging task because music itself is an art form full of complex emotions and subjectivity. The expression of emotions varies greatly between different musical compositions, resulting in models that may require more fine-tuning and dataset diversity in accurately understanding and interpreting these emotions. In addition, the fusion and co-processing of multimodal data is a complex area that requires further research and improvement. In our model, the fusion of visual and auditory information provides unique diversity and creativity in the art of robotic musical performance, but more work is still needed on how to integrate this information more effectively to improve the performance and robustness of the model. This includes more advanced fusion algorithms, larger multimodal datasets, and a

TABLE 5 Ablation experiments on the GoogLeNet module using different datasets.

Model	Datasets							
	Music Emotion Dataset				Multimodal DataSet			
	Accuracy	Recall	F1 Score	AUC	Accuracy	Recall	F1 Score	AUC
ResNet	93.75	88.89	87.7	90.39	89.47	86.9	85.74	90.29
VGGNet	94	93.38	84.57	86.38	87.08	87.73	84.8	88.74
DenseNet	95.9	86.73	90.68	84.74	95.07	93.53	86.95	90.8
GoogLeNet	96.98	93.46	92.67	93.22	96.98	93.46	92.67	93.22

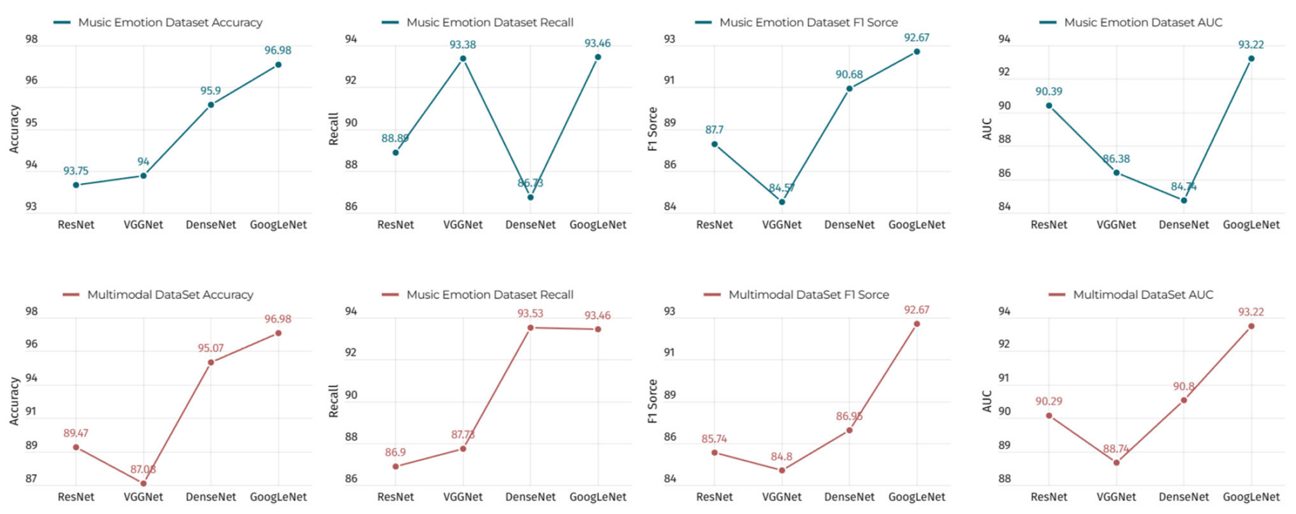

FIGURE 11
Comparison of the performance of the GoogLeNet model and other models under different conditions.

deeper understanding of the complementarities between different modalities. Meanwhile, the accuracy of complex emotional states and different audience interactions still needs to be improved.

In our next research, we plan to further deepen our music sentiment analysis techniques to improve the model's ability to understand and express musical emotions. This will involve more refined algorithm development and more extensive data acquisition to capture and interpret the complex and subtle emotional layers in musical compositions. At the same time, we will continue to investigate methods for fusing multimodal data to achieve a more comprehensive audiovisual perception. Our goal is to create a system that seamlessly integrates visual and auditory information so that the robot can more accurately interpret and respond to audience emotions and interactions. By doing so, we hope to improve the overall performance of the model, especially in dealing with complex emotional states and diverse audience interactions. A key focus of future work will be to refine the model's emotional intelligence and its adaptability to different performance contexts. This includes adapting the model to better understand and reflect emotional expressions in different cultural and social contexts. In addition, we will explore the integration of augmented reality elements to further enrich the audience experience. By integrating augmented reality into our system, we can provide audiences with more diverse and interactive performances, adding innovative elements on both the visual and perceptual levels. The research in this paper has important implications for the field of robotic music performance art and multimodal data processing. Our research provides innovative ideas and methods for multidisciplinary applications, and we expect that this research will advance the development of music and multimodal technologies and open up new possibilities for future robotic music performance art.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

LW: Conceptualization, Formal analysis, Funding acquisition, Methodology, Resources, Software, Writing - original draft, Writing - review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Andersson, U., Cuervo-Cazurra, A., and Nielsen, B. B. (2020). "Explaining interaction effects within and across levels of analysis," in *Research Methods in International Business* (Berlin: Journal of International Business Studies), 331–349.

Bhat, A., and Jain, S. (2023). Face recognition in the age of clip & billion image datasets. *arXiv preprint arXiv:2301.07315*. doi: 10.48550/arXiv.2301.07315

Erickson, Z., Xing, E., Srirangam, B., Chernova, S., and Kemp, C. C. (2020). "Multimodal material classification for robots using spectroscopy and high resolution texture imaging," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Piscataway, NJ: IEEE), 10452–10459.

Funk, M., Kuwabara, K., and Lyons, M. J. (2020). Sonification of facial actions for musical expression. *arXiv preprint arXiv:2010.03223*. doi: 10.5281/zenodo.1176749

Gao, T., Wang, C., Zheng, J., Wu, G., Ning, X., Bai, X., et al. (2023). A smoothing group lasso based interval type-2 fuzzy neural network for simultaneous feature selection and system identification. *Knowledge Based Syst.* 111028. doi: 10.1016/j.knosys.2023.111028

Ghaleb, E., Niehues, J., and Asteriadis, S. (2020). "Multimodal attention-mechanism for temporal emotion recognition," in *2020 IEEE International Conference on Image Processing (ICIP)* (Piscataway, NJ), 251–255.

Gouyon, F., Dixon, S., Pampalk, E., and Widmer, G. (2006). "Evaluating rhythmic descriptors for musical genre classification," in *ISMIR* (London), Vol. 6, 309–314.

Hariri, W. (2022). Efficient masked face recognition method during the covid-19 pandemic. *Signal Image Video Process.* 16, 605–612. doi: 10.1007/s11760-021-02050-w

Inamura, T., and Mizuchi, Y. (2021). SIGverse: a cloud-based VR platform for research on multimodal human-robot interaction. *Front. Robot. AI* 8, 549360. doi: 10.3389/frobt.2021.549360

Koelsch, S., Fritz, T., Cramon, D. Y., Meller, K., and Friederici, A. D. (2006). Investigating emotion with music: an fMRI study. *Hum. Brain Mapp.* 27, 239–250. doi: 10.1002/hbm.20180

Laghrissi, F., Douzi, S., Douzi, K., and Hssina, B. (2021). Intrusion detection systems using long short-term memory (LSTM). *J. Big Data* 8, 65. doi: 10.1186/s40537-021-00448-4

Lee, C.-H., Liu, Z., Wu, L., and Luo, P. (2020). "MaskGAN: towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Piscataway, NJ), 5549–5558.

Lee, H. (2019). Real-time manufacturing modeling and simulation framework using augmented reality and stochastic network analysis. *Virt. Real.* 23, 85–99. doi: 10.1007/s10055-018-0343-6

Li, H. (2022). *Learning to Rank for Information Retrieval and Natural Language Processing*. Berlin: Springer Nature.

Li, L., Mu, X., Li, S., and Peng, H. (2020). A review of face recognition technology. *IEEE Access* 8, 139110–139120. doi: 10.1109/ACCESS.2020.3011028

Li, T., and Ogihara, M. (2003). "Detecting emotion in music," in *4th International Conference on Music Information Retrieval* (Baltimore, MD).

Li, X., Ren, J., and Li, D. (2023). Hierarchical matching algorithm of visual image for mobile robots based on deep reinforcement learning. *J. Jilin Univ. Sci. Ed.* 61, 127–135.

Liu, Q., Zheng, C., He, H., and Wu, L. (2022). A variable step size lms speech denoising algorithm based on wavelet threshold. *J. Jilin Univ. Sci. Ed.* 60, 943–949.

Müller, M., Kurth, F., and Clausen, M. (2005). "Audio matching via chroma-based statistical features," in *ISMIR*, Vol. 2005 (London: Citeseer), 6.

Ning, X., Tian, W., He, F., Bai, X., Sun, L., and Li, W. (2023). Hyper-sausage coverage function neuron model and learning algorithm for image classification. *Pattern Recogn.* 136, 109216. doi: 10.1016/j.patcog.2022.109216

Pandeya, Y. R., Bhattarai, B., and Lee, J. (2021). Deep-learning-based multimodal emotion classification for music videos. *Sensors* 21, 4927. doi: 10.3390/s21144927

Pang, Y., Zhao, X., Zhang, L., and Lu, H. (2020). "Multi-scale interactive network for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Piscataway, NJ), 9413–9422.

Pepino, L., Riera, P., and Ferrer, L. (2021). Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502.*

Petrović, N. (2020). "Augmented and virtual reality web applications for music stage performance," in *2020 55th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)* (Piscataway, NJ), 33–36.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., and Huang, X. (2020). Pre-trained models for natural language processing: a survey. *Sci. China Technol. Sci.* 63, 1872–1897. doi: 10.1007/s11431-020-1647-3

Schuller, B. W., Valstar, M., Eyben, F., Cowie, R., and Pantic, M. (2010). Mir_kovic: a multimodal database infrastructure for research on audio-visual affective expression. *IEEE Trans. Affect. Comput.* 1, 83–98.

Selvi, R., and Vijayakumaran, C. (2023). "Stocastic multimodal fusion method for classifying emotions with attention mechanism using deep learning," in *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)* (Piscataway, NJ), Vol. 1, 2347–2352.

Staudemeyer, R. C., and Morris, E. R. (2019). Understanding LSTM–a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586*. doi: 10.48550/arXiv.1909.09586

Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *J. Acoust. Soc. Am.* 8, 185–190.

Sun, G., Wong, Y., Cheng, Z., Kankanhalli, M. S., Geng, W., and Li, X. (2020). Deepdance: music-to-dance motion choreography with adversarial learning. *IEEE Trans. Multimedia* 23, 497–509. doi: 10.1109/TMM.2020.2981989

Swarbrick, D., Bosnyak, D., Livingstone, S. R., Bansal, J., Marsh-Rollo, S., Woolhouse, M. H., et al. (2019). How live music moves us: head movement differences in audiences to live versus recorded music. *Front. Psychol.* 9, 2682. doi: 10.3389/fpsyg.2018.02682

Tang, Q., Huang, Z., Zhou, H., and Ye, P. (2020). Effects of music therapy on depression: a meta-analysis of randomized controlled trials. *PLoS ONE* 15, e0240862. doi: 10.1371/journal.pone.0240862

Tsiourti, C., Weiss, A., Wac, K., and Vincze, M. (2019). Multimodal integration of emotional signals from voice, body, and context: effects of (in) congruence on emotion recognition and attitudes towards robots. *Int. J. Soc. Robot.* 11, 555–573. doi: 10.1007/s12369-019-00524-z

Vakalopoulou, M., Christodoulidis, S., Sahasrabudhe, M., Mougiakakou, S., and Paragios, N. (2019). "Image registration of satellite imagery with deep convolutional neural networks," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium* (Piscataway, NJ), 4939–4942.

Wang, W., Li, R., Chen, Y., Sun, Y., and Jia, Y. (2021). Predicting human intentions in human–robot hand-over tasks through multimodal learning. *IEEE Trans. Automat. Sci. Eng.* 19, 2339–2353. doi: 10.1109/TASE.2021.3074873

Yu, Z., Dong, Y., Cheng, J., Sun, M., and Su, F. (2022). Research on face recognition classification based on improved googlenet. *Sec. Commun. Netw.* 2022, 1–6. doi: 10.1155/2022/7192306

Zadeh, A., Liang, P. P., Poria, S., and Cambria, E. (2018). "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Florence), 361–372.

Zhang, S. (2020). The positive influence of music on the human brain. *Behav. Brain Sci.* 10, 95–104. doi: 10.4236/jbbs.2020.101005

Zhou, H., Zhang, Y., Yang, L., Liu, Q., Yan, K., and Du, Y. (2019). Short-term photovoltaic power forecasting based on long short term memory neural network and attention mechanism. *IEEE Access* 7, 78063–78074. doi: 10.1109/ACCESS.2019.2923006

Zhu, C., Cheng, Y., Gan, Z., Sun, S., Goldstein, T., and Liu, J. (2019). FreeLB: enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*. doi: 10.48550/arXiv.1909.11764