Employee Emotion Recognition and Job Satisfaction Enhancement in Hotel Management: A Convolutional Neural Network-Based Solution

Babar Shahzad*

Department of Management Sciences, COMSATS University Islamabad; babarshehzad2345@gmail.com

*Corresponding Author: babarshehzad2345@gmail.com

DOI: https://doi.org/10.30210/JMSO.202604.009

Received: Feb 23, 2026 Accepted: May 05, 2026

ABSTRACT

As the impact of employees' emotions on job satisfaction and performance has been increasingly emphasized, how to accurately predict employees' job satisfaction has become an important topic in the management field. However, existing satisfaction prediction methods usually rely on a single data source and fail to fully consider the dynamic changes and diversity of emotions, resulting in less accurate prediction results. To address this issue, this study proposes the EmoFusionNet model, which combines emotion recognition with job satisfaction prediction by fusing facial expression images and speech emotion signals, taking full advantage of the complementary nature of multimodal data. The model uses feature-weighted fusion, emotion classification and regression analysis to effectively improve the accuracy of employee job satisfaction prediction. Experimental results show that EmoFusionNet significantly outperforms traditional unimodal methods in multimodal emotion recognition tasks, especially in employee emotion recognition and real-time satisfaction prediction, exhibiting high accuracy and robustness. This study provides an intelligent emotion management tool for industries such as hotel management, and also provides a useful reference for future work environment optimization based on multimodal data.

Keywords: Multimodal emotion recognition, Vision transformer, Temporal convolutional networks, Job satisfaction prediction, Emotion classification, Deep learning

1. Introduction

With the rapid development of artificial intelligence technology, emotion recognition technology has become a key research direction in several fields, especially in the hospitality industry, where employee's emotion directly affects customer satisfaction, service quality, and employee work efficiency[1]. The emotional state of employees not only determines their work performance, but also has a profound impact on the teamwork atmosphere and overall job satisfaction[2]. Therefore, how to accurately identify employees' emotions in real time, especially in high-pressure work environments, has become a key issue to optimize the work environment and improve employee satisfaction.

Traditional emotion recognition techniques mostly rely on a single data source, such as facial expressions or speech signals, and usually classify emotions through deep learning models such as convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) [3]. Although these methods have achieved some results in certain scenarios, they often show some limitations in multimodal data processing and capturing long time dependencies[4] [5]. Especially when facing complex emotional states, it is difficult for a single data source to comprehensively capture the multidimensional features of emotions, resulting in limited accuracy and effectiveness of emotion recognition.

In recent years, research based on novel deep learning architectures such as Vision Transformer (ViT) and Temporal Convolutional Networks (TCN) has gained widespread attention.ViT efficiently captures global information in image processing through the mechanism of self-attention, while TCN utilizes convolutional operations to efficiently extract dynamic features from time series data. dynamic features in time-series data using convolutional operations. This makes these architectures show more advantages in multimodal emotion recognition tasks, especially in high-dimensional data processing and capturing emotion changes.

In this context, combining multimodal data for emotion recognition has become an effective way to improve the accuracy and comprehensiveness of emotion recognition. Unlike traditional single data source models, combining facial expressions, speech signals, and physiological data (e.g., heart rate and EEG, etc.) can more comprehensively and accurately capture employees' emotional changes. Through the collaborative work of these multimodal data, it can not only improve the accuracy of emotion recognition, but also provide a strong data support to enhance employee job satisfaction.

Based on this research background, this study proposes EmoFusionNet, a multimodal emotion recognition model that combines Vision Transformer (ViT) and Temporal Convolutional Networks (TCN). The model significantly improves the accuracy of emotion recognition by processing multiple data sources such as facial expressions and speech signals, and predicts employees' job satisfaction by combining emotional states. The contributions of this paper are mainly in the following aspects:

- Proposal multimodal emotion recognition model: by combining ViT and TCN, an efficient multimodal emotion recognition model is proposed, which is able to process multiple data such as facial expression images and speech signals at the same time to improve the accuracy of emotion recognition.
- Emotion and job satisfaction correlation analysis: based on the emotion recognition results, this paper explores the relationship between employee emotions and job satisfaction, and predicts employee job satisfaction through deep learning model to provide decision support for hotel management.
- Model Performance Evaluation and Comparison: Through a large number of experiments, this paper verifies the advantages of EmoFusionNet in emotion recognition accuracy and job satisfaction prediction, and compares it with the traditional CNN+LSTM model, demonstrating the excellent performance of the new model in complex tasks.

This paper is organized as follows: Chapter 2 will present a relevant literature review and review

the current major techniques and methods in the field of emotion recognition; Chapter 3 describes in detail the design of the EmoFusionNet model, including the data preprocessing, feature extraction, emotion recognition and job satisfaction prediction modules; Chapter 4 presents the experimental design and the result analysis, and discusses the performance of the model as well as the comparison with the existing methods; Finally, Chapter 5 summarizes the research results of this paper and looks forward to future research directions.

2. Literature Review

Emotion recognition technology has been widely used in many fields, especially in the hotel industry, where employee emotions have a significant impact on service quality and customer satisfaction. Emotion recognition research mainly focuses on the processing of data such as facial expressions, voice, body language, and physiological signals. With the continuous development of deep learning technology, the accuracy and application breadth of emotion recognition are also constantly improving. The following will review the research progress in this field and summarize the current challenges.

2.1 Traditional Emotion Recognition Methods

Early emotion recognition methods mainly rely on traditional machine learning algorithms, such as support vector machines (SVM), random forests (RF), and k-nearest neighbors (KNN). Most of these methods rely on artificial feature extraction[6], such as facial expression features (using Haar cascade feature detectors) or speech signal features (such as MFCC) for classification[7] [8]. However, these methods are highly dependent on data and are difficult to adapt to complex and changeable emotional expressions[9] [10]. In particular, when faced with emotional diversity and individual differences, traditional methods are difficult to achieve efficient and accurate emotion classification[11].

With the introduction of deep learning technology, convolutional neural networks (CNN) and long short-term memory networks (LSTM) have become the mainstream methods for emotion recognition[12] [13]. CNN has made significant progress in facial expression image processing, and can automatically extract features from images, avoiding the shortcomings of manual feature extraction. LSTM has been widely used in emotion recognition of speech and time series data, especially it can effectively capture the dynamic characteristics of emotions changing over time[14], improving the accuracy of emotion recognition[15].

2.2 Research on Multimodal Emotion Recognition

In recent years, the research on emotion recognition has gradually developed in the direction of multimodality[16]. Traditional emotion recognition from a single data source (such as facial expressions or speech) often has certain limitations[17] [18], especially in terms of the diversity and complexity of emotions. Multimodal emotion recognition methods can provide more comprehensive and accurate emotion judgments by combining data from different sources such as facial expressions[19], voice signals, and physiological signals. Multimodal emotion recognition can not

only make up for the shortcomings of a single data source[20], but also improve the robustness of the model.

At present, the main challenge of multimodal emotion recognition is how to effectively fuse data from different modalities. Researchers have proposed a variety of fusion strategies, including directly splicing the features of each modality and fusing features by weighted average[21]. Although these methods can improve recognition accuracy, how to solve the heterogeneity of different modal data and how to deal with noise in the data during data fusion are still important research topics in this field.

2.3 Application and Challenges of Deep Learning Models

Deep learning models, especially convolutional neural networks (CNNs) and long short-term memory networks (LSTMs), have achieved remarkable results in emotion recognition[22]. However, with the deepening of research, many new architectures have begun to attract the attention of researchers, such as self-attention mechanisms and graph neural networks (GNNs). These new models have strong feature extraction capabilities and time series modeling capabilities, and can play a greater role in processing multimodal emotion data[23].

The self-attention mechanism can help the model focus on important moments when processing time series data and improve the accuracy of emotion recognition. Graph neural networks can better handle complex data structures[24], especially when processing physiological signals or emotion-related social interaction data. GNNs can effectively model the relationship between data. Although these new models have great potential[25], there are still many challenges in how to efficiently apply them to multimodal emotion recognition tasks, especially when facing high-dimensional and heterogeneous data.

In addition, the efficiency and real-time performance of emotion recognition technology in practical applications are also important research directions[26]. In a commercial environment, emotion recognition systems need to process a large amount of real-time data. How to improve computing speed and reduce latency while ensuring accuracy is a key issue in current research.

3. Method

3.1 Model Design and Methods

At present, one of the main problems in the field of emotion recognition is how to effectively process multimodal data and accurately classify emotions. Traditional emotion recognition methods usually rely on a single data source, such as facial expressions or voice signals. This method cannot fully capture the multidimensional characteristics of emotions, resulting in low accuracy in emotion classification. In addition, although deep learning methods such as convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) have achieved some success in emotion recognition, they still have certain limitations in processing multimodal data and long-term dependent emotion changes. Therefore, how to improve the efficiency of multimodal data fusion, capture the dynamic changes of emotions, and provide accurate job satisfaction predictions on this basis remains a key challenge in this field.

In order to overcome these problems, based on previous studies, this paper draws on the advantages of multimodal emotion recognition and deep learning models, and proposes a multimodal emotion recognition model based on EmoFusionNet. This model combines facial expression images and voice signals, uses a deep learning network to accurately classify emotions, and predicts employee job satisfaction through emotion recognition results. Compared with the traditional single data source method, EmoFusionNet can capture emotional information more comprehensively and has better performance in handling long-term dependencies and multimodal data fusion.

This model consists of three main parts: feature extraction module, emotion recognition module and job satisfaction prediction module. The function of the feature extraction module is to extract features that can effectively express emotions from facial expression images and voice signals; the emotion recognition module processes the extracted features to judge the emotional state of employees; the job satisfaction prediction module evaluates the job satisfaction of employees through the emotion recognition results. These three parts together constitute a complete emotion recognition and satisfaction prediction system.

First, after the data is input into the model, the facial expression images and voice signals will be extracted through two different networks respectively. In the feature extraction of facial expressions, convolutional neural networks (CNNs) are used to automatically extract emotion-related features from images. CNN obtains key information in images from simple to complex through the gradual extraction of multiple convolutional layers, which can effectively identify employees' facial expressions and reflect their emotional states. At the same time, the emotional features of the speech signal are extracted through Mel Frequency Cepstral Coefficients (MFCC), which effectively captures the pitch, rhythm and timbre changes in the speech, providing rich acoustic information for emotion recognition.

Next, the extracted facial expression features and speech signal features will be fed into the emotion recognition module. This module fuses and classifies the data of two different modalities through a deep neural network. The core task of the emotion recognition module is to judge the current emotional state of the employee based on the characteristics of facial expressions and speech signals. The classification results will include different emotion categories, such as happiness, anxiety, anger, etc.

Finally, the results of emotion recognition will be passed as input to the job satisfaction prediction module. This module predicts the job satisfaction of employees based on their emotional state and historical data. The job satisfaction prediction module uses a regression model to score job satisfaction, and helps managers make appropriate decisions through the relationship between emotional state and satisfaction.

The advantage of this model lies in its efficient processing and fusion of multimodal data, which can extract emotional features from facial expressions and speech signals at the same time, and improve the accuracy of emotion recognition. Compared with the traditional single data source model, EmoFusionNet can more comprehensively understand the emotional state of employees through the multi-dimensional capture of emotions. At the same time, this model can also predict employees' job

satisfaction through their emotional state, provide effective decision-making support for enterprises, and help managers optimize employees' working environment and management strategies.

In terms of expected results, EmoFusionNet is expected to provide high accuracy in emotion recognition tasks and have advantages in real-time and computational efficiency. By integrating multimodal data, the model can better adapt to the actual needs of the hotel industry, especially in high-pressure and complex working environments, and monitor employee emotional changes in real time, thereby improving job satisfaction and service quality.

3.2 Feature Fusion and Emotion Classification

In the EmoFusionNet model, the feature fusion and emotion classification module is the core part of the model. Its main task is to effectively fuse the features extracted from the facial expression image and the voice signal, and classify emotions based on the fused features. Feature fusion can not only integrate the information of the two modalities, but also help the model to understand the emotional state of employees more comprehensively, thereby improving the accuracy of emotion classification.

Facial expressions and voice signals provide different aspects of emotions. Facial expression images mainly reflect the immediate emotional reactions of employees, while voice signals can reveal information about the tone, rhythm, and pitch of emotions. Therefore, before emotion classification, it is necessary to fuse the features from different modalities so that the model can take advantage of the advantages of both modalities at the same time.

Assume that the features extracted from the facial expression image and the voice signal are F and V, respectively, where F is the feature vector extracted from the facial expression image and V is the feature vector extracted from the voice signal. In order to effectively fuse the two feature vectors, we use the weighted sum method to obtain the fused feature Z. The weighted sum formula is as follows:

$$Z = \alpha F + \beta V \quad \text{[Formular 1]}$$

where α and β are the weights assigned to the face image features and the audio features, respectively, and Z is the fused feature vector.

Through weighted fusion, we are able to balance the impact of facial expressions and voice signals on emotions. Different emotional expressions may have different weights in facial expressions and voices, so by learning the best α and β values, the model is able to optimize the representation of fused features.

The goal of the emotion classification module is to judge the emotional state of employees based on the fused feature Z. To achieve emotion classification, the model maps the fused features to different emotion categories through a multi-layer fully connected network (also called a fully connected layer). Emotion categories may include happiness, anxiety, fatigue, anger, etc. The calculation process of emotion classification can be used to predict the probability distribution through the Softmax function.

The output C of emotion classification can be expressed by the following formula:

$$C = \text{softmax}(WZ + b) \quad \text{[Formular 2]}$$

where W is the weight matrix, b is the bias vector, and C is the output vector representing the probabilities of different emotion classes.

The role of the Softmax function is to convert the linear combination of the model output (i.e. $WZ + b$) into a probability distribution, thereby representing the probability of each emotion category. Finally, the model determines the predicted emotion category based on the maximum probability.

In the process of emotion classification, the goal of the model is to minimize the classification error, that is, to optimize the accuracy of emotion classification by continuously adjusting the weight W and bias b. To this end, the model uses the cross-entropy loss function to measure the difference between the predicted emotion category and the true label. The cross entropy loss function can be expressed as:

$$L_{\text{CE}} = -\sum_{i=1}^{N} y_i \log(p_i) \quad \text{[Formular 3]}$$

where N is the number of classes, y_i is the ground truth label for the i-th class, and p_i is the predicted probability for the i-th class.

By minimizing the cross entropy loss function, the model can optimize the accuracy of emotion classification. During the training process, the Adam optimizer is used to dynamically adjust the learning rate and accelerate the convergence of the model. To help visualize the architecture of the feature fusion and emotion classification process in EmoFusionNet, the following Figure 1 illustrates the core structure of the model. This diagram highlights the role of the Vision Transformer (ViT) for extracting facial expression features, the Temporal Convolutional Networks (TCN) for processing voice signals, and the Self-Attention mechanism that enhances the model's focus on key emotional features.

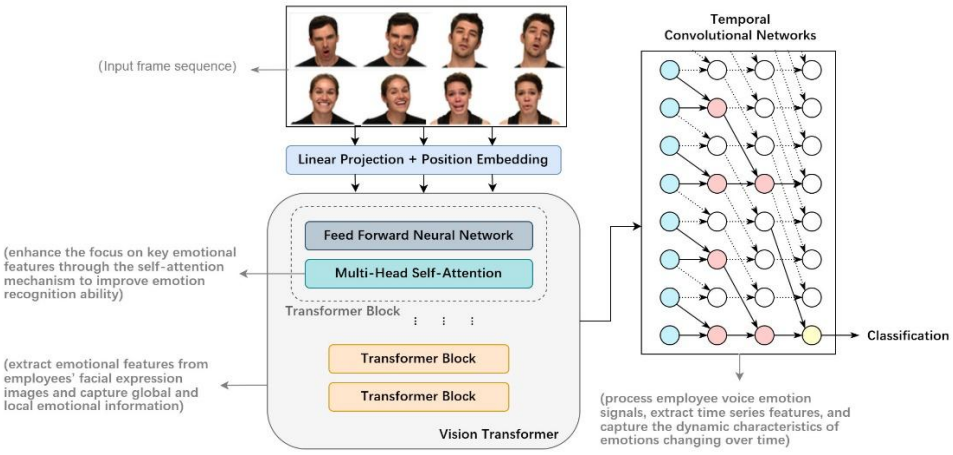

Figure 1. Architecture of emofusionnet: feature fusion and emotion classification

Through the above-mentioned feature fusion and emotion classification design, the EmoFusionNet model has significant advantages. First, multimodal data fusion can make up for the limitations of a single data source. Facial expression images and voice signals are the two main ways to express emotions. Combining the features of these two modalities, the model can capture emotional

information more accurately. Secondly, the emotion classification module can automatically learn the emotion patterns in facial expressions and voice signals through the multi-layer structure of the deep neural network, thereby improving the accuracy of classification.

In addition, using the Softmax function for emotion classification allows the model to output the probability distribution of each emotion category, enhancing the interpretability and operability of emotion prediction. Based on this structure, EmoFusionNet shows strong robustness and accuracy when processing complex emotional expressions, and can provide more reliable emotion recognition and satisfaction prediction in practical applications.

3.3 Job Satisfaction Prediction

In the EmoFusionNet model, the job satisfaction prediction module is a key module closely connected with the emotion recognition module. Its main task is to predict the job satisfaction of employees based on their emotional state. There is a close relationship between emotions and job satisfaction, so accurately understanding the emotional state of employees helps to better evaluate their job satisfaction. This section will introduce the design and implementation of the job satisfaction prediction module in detail, focusing on how to use emotion recognition results to predict satisfaction scores.

Job satisfaction is an employee's psychological response to their work environment, work content, work interpersonal relationships, and overall work experience. Studies have shown that emotions play a vital role in employees' job satisfaction. For example, employees usually show higher job satisfaction when they are happy and positive; while in the case of anxiety and excessive stress, employees may be dissatisfied with their work, thereby reducing overall satisfaction. Therefore, accurate emotion recognition is the basis for job satisfaction prediction.

In this model, the emotion classification module will output the current emotional state of employees, and the emotion categories may include happiness, anxiety, fatigue, etc. The job satisfaction prediction module will predict the employee's job satisfaction score based on these emotion classification results. The model combines the emotion classification results with historical data through regression analysis and outputs a continuous satisfaction score M, which represents the employee's job satisfaction level.

The calculation of the work satisfaction score M depends on the emotion classification result C. The model will perform weighted processing according to different emotion categories, and combine the individual differences of employees and historical data to finally output the work satisfaction score. In this process, we use regression methods (such as linear regression or multi-layer perceptron) to map the relationship between emotional state and job satisfaction.

Assuming that the emotion classification result is C, the work satisfaction score M can be expressed by the following formula:

$$M = \mathrm{ReLU}(W_c C + b_c) \quad \text{[Formular 4]}$$

where W_c is the weight matrix for the emotional classification output, b_c is the bias term, and M is the predicted work satisfaction score. The ReLU function is used to ensure that the predicted

satisfaction score is non-negative.

Through this regression method, the model can estimate the employee's job satisfaction based on his or her emotional state. This process takes into account the impact of emotions on job satisfaction and outputs a real-time satisfaction score based on the employee's emotional state.

The training and optimization process of the job satisfaction prediction module is similar to that of the emotion classification module. In order to optimize the prediction accuracy of job satisfaction, the model uses mean square error (MSE) as the loss function to calculate the difference between the predicted value and the true satisfaction score. The mean square error formula is as follows:

$$L_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(M_i - \widehat{M}_i)^2 \qquad \text{[Formular 5]}$$

where N is the number of samples, M_i is the true satisfaction score, and $\widehat{M}_i$ is the predicted satisfaction score. By minimizing the mean square error, the model can continuously adjust parameters to make job satisfaction prediction more accurate. During the training process, the Adam optimizer is used to accelerate convergence and dynamically adjust the learning rate to improve training efficiency.

To better visualize the architecture and performance of the job satisfaction prediction process, the following Figure 2 demonstrates how the model employs Support Vector Machine (SVM) for the classification of emotional data, which plays a critical role in predicting job satisfaction. The left side of the figure shows the data in its original form, where the red and blue points are not linearly separable by a simple straight line in the 2D feature space. This scenario is typical when emotions are not easily separated by simple decision boundaries.

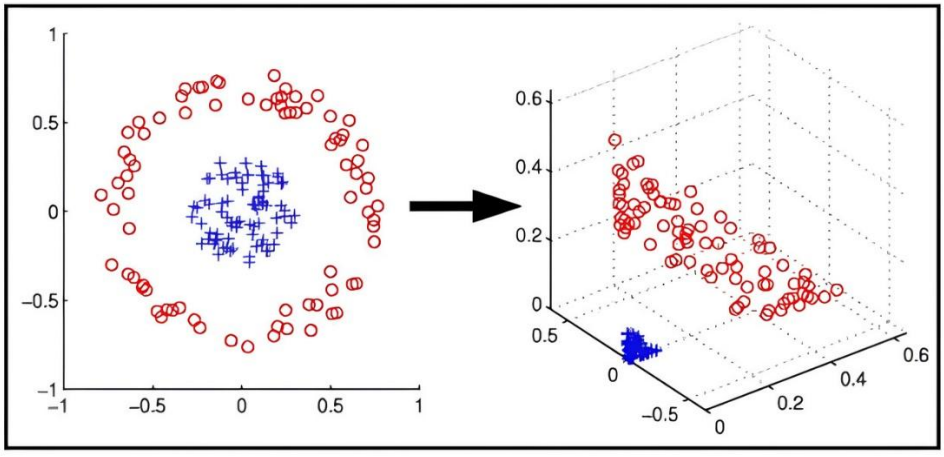

Figure 3. SVM Application: mapping non-linearly separable data to higher dimensions

However, as shown in the right side of the figure, SVM uses a technique called kernel trick to map the original data into a higher-dimensional space (3D in this case), where the data points become linearly separable. By doing so, SVM can effectively find a hyperplane (represented by the separating plane) that maximizes the margin between the red and blue classes, allowing the model to classify emotional states more accurately.

This process of mapping the data into a higher-dimensional space and finding the optimal hyperplane directly impacts the prediction of job satisfaction. The ability to separate emotional states effectively in this higher-dimensional space enables the model to make more accurate predictions of how employees' emotional states influence their job satisfaction.

The job satisfaction prediction module of EmoFusionNet has significant advantages over traditional satisfaction prediction methods. First, the model accurately captures the emotional changes of employees through emotion recognition results, and models the relationship between these changes and job satisfaction, which can reflect the work status of employees in real time. Second, the job satisfaction score is a continuous value. The model not only gives an emotion classification, but also provides a more detailed and quantitative satisfaction score. This enables the model to provide managers with more comprehensive feedback, thereby helping to improve the work experience of employees and the overall effectiveness of the team.

In terms of expected effects, EmoFusionNet is expected to have high accuracy in emotion recognition and job satisfaction prediction, which can help hotel managers better understand employee emotions and improve job satisfaction. Through this intelligent emotion management tool, employee satisfaction can be monitored and optimized in real time, thereby improving the overall work performance and service quality of employees.

4. Experiment

4.1 Data Collection and Preprocessing

This study used multiple datasets to train and verify the EmoFusionNet model. The main data sources include employees' facial expression images, voice emotion signals, and physiological signals. Facial expression images and voice emotion signals mainly come from public datasets, such as the EmoReact dataset[26] and the RAVDESS dataset[27]. These two datasets contain facial expression images and voice signals in a variety of emotional states, covering emotion categories such as happiness, anxiety, anger, and sadness. Each data point has a clear emotion label, providing sufficient training samples for the emotion recognition model. In addition, this study also collected data from actual work environments, including employees' heart rate and EEG signals, in order to further improve the accuracy of emotion recognition and job satisfaction prediction.

In the data preprocessing stage, facial expression images and voice signals need to be processed first. When preprocessing image data, the size is resized and the pixel values are normalized. All facial expression images are resized to a uniform size (for example: 224x224 or 256x256 pixels) to ensure that the network can process images of different resolutions. The pixel value normalization operation scales the pixel values of the image from the range of [0, 255] to [0, 1] or [-1, 1] to avoid training instability due to data range differences. In addition, in order to enhance data diversity, the image data is augmented, including rotation, flipping, scaling, and random cropping. These enhancement operations help the model better adapt to facial expression changes in various real-world environments.

The processing of speech signals includes feature extraction and normalization of audio signals.

We use the Mel Frequency Cepstral Coefficient (MFCC) method to extract the features of audio signals, which can effectively capture the emotion-related information such as timbre, pitch, and rhythm in speech. In the preprocessing of audio data, the audio is first framed and windowed to ensure that the time domain information of the audio signal is fully represented. Usually, we extract 13-dimensional MFCC coefficients for each audio segment, and the length of each frame is 25 milliseconds and the frame shift is 10 milliseconds. Such processing can ensure that the emotional information of the audio signal is accurately captured.

For physiological signal data, including employees' heart rate and EEG signals, the preprocessing steps mainly include normalization and filtering. Heart rate signals are usually recorded in the form of beats per minute (bpm), and the model needs to normalize them to the range of [0, 1] to be consistent with other input features. EEG signals need to be processed through a bandpass filter to remove unnecessary high-frequency noise and low-frequency drift. The filtered EEG signals can provide employees' brain electrical activity data, which helps analyze emotional states. All these physiological signals are standardized and can be input into the model as supplementary information to further improve the effect of emotion recognition and job satisfaction prediction.

In order to improve the generalization and adaptability of the model, data enhancement and sample balancing techniques are also applied. In image data processing, through data enhancement methods such as rotation, translation, cropping and color adjustment, the model can see more diverse emotional expressions, thereby enhancing its adaptability to emotional changes. In terms of speech data enhancement, by adjusting the volume, speed and background noise of the speech, the model can adapt to different audio environments, which is of great significance for speech signals in practical applications.

In terms of sample balance, some emotion categories have fewer samples, which may lead to category imbalance. In order to balance the dataset, this paper uses oversampling and undersampling techniques. Oversampling balances the category distribution in the dataset by increasing the number of samples in the minority category; undersampling avoids overfitting by reducing the number of samples in the majority category. These methods ensure that the model can effectively learn the characteristics of each emotion category during training, thereby improving the accuracy of emotion classification.

By preprocessing facial expression images, voice signals, and physiological signal data, this study ensures that multimodal data can be input into the model in a unified format, and enhances the generalization ability and stability of the model through data enhancement and sample balancing techniques. These preprocessing steps lay a solid foundation for subsequent feature extraction and emotion classification, thereby improving the overall performance of emotion recognition and job satisfaction prediction.

4.2. Experimental Setup

In this study, the experimental design includes dataset division, evaluation criteria, and hardware and software environment. The experimental setup provides a reliable basis for the training and evaluation of the EmoFusionNet model.

4.2.1 Dataset division and evaluation criteria

In order to evaluate the performance of the EmoFusionNet model, we divide the dataset into training set, validation set, and test set. The training set accounts for 70% of the dataset, the validation set accounts for 15%, and the test set accounts for 15%. This division method ensures that the model can learn sufficient samples during training, while hyperparameters are adjusted through the validation set, and the generalization ability of the model is finally evaluated through the test set.

In terms of evaluation criteria, in addition to the conventional accuracy, this study also uses the weighted average accuracy (Weighted Accuracy) and AUC (Area Under the Curve) to evaluate the performance of the model. The weighted average accuracy takes into account the problem of class imbalance. It calculates the accuracy by weighting the number of samples in each class. The formula is as follows:

$$\text{Weighted Accuracy} = \sum_{i=1}^{N} \frac{n_i}{N} \cdot \text{Accuracy}_i \quad \text{[Formular 6]}$$

where n_i is the number of samples in the i -th class, N is the total number of samples, and Accuracy_i is the accuracy for the i-th class.

AUC (area under the curve) is an important indicator for measuring the performance of a classification model. It indicates the comprehensive performance of the model under different classification thresholds. The closer the AUC is to 1, the better the model performance. The AUC calculation formula is as follows:

$$\text{AUC} = \int_0^1 \text{True Positive Rate}(t)\,, \text{dFalse Positive Rate}(t) \quad \text{[Formular 7]}$$

In addition, for job satisfaction prediction, the root mean square error (RMSE) is used as the evaluation indicator. RMSE can measure the average error of the model in predicting job satisfaction, and the formula is as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(M_i - \widehat{M}_i)^2} \quad \text{[Formular 8]}$$

where M_i is the true satisfaction score and $\widehat{M}_i$ is the predicted satisfaction score.

4.2.2 Hardware and software environment

High-performance hardware and software environment are used in the experiment to ensure the efficiency and scalability of the training process. The specific configuration is as follow Table 1:

Table 1. Hardware and software environment configuration

Item	Description
Hardware	**NVIDIA Tesla V100 GPU** for deep learning acceleration
Processor	Intel Xeon CPU (16 cores)
Memory	64GB DDR4 RAM
Storage	1TB SSD
Software	Python 3.8, PyTorch 1.9

Item	Description
Operating System	Ubuntu 20.04
GPU Acceleration	NVIDIA CUDA 11.2, cuDNN
Frameworks/Libraries	NumPy, Matplotlib, Seaborn, Scikit-learn
Optimizer	Adam Optimizer
Training Epochs	100 epochs
Batch Size	64

4.2.3 Training and tuning

During the training process, the Adam optimizer is used to optimize the weight parameters of the model. The Adam optimizer has the ability to dynamically adjust the learning rate, which can accelerate the convergence speed of the model and avoid overfitting. The training rounds for each experiment are set to 100 rounds, and the batch size is 64 to ensure that the model can process enough samples in each training. After each round of training, the model is evaluated on the validation set, and the hyperparameters are adjusted according to the performance of the validation set.

In order to further improve the performance of the model, the cross-validation method is used in the experiment to tune the model, and L2 regularization is introduced to prevent overfitting. By monitoring the evaluation indicators during the training process, it is ensured that the model performs best in both emotion recognition and job satisfaction prediction tasks.

4.3. Experimental Results and Analysis

4.3.1 Performance comparison experiment

Table 2 shows multiple evaluation indicators of the EmoFusionNet model on the training set, validation set, and test set, including accuracy, AUC, weighted accuracy, and macro F1 score. These indicators can fully reflect the performance of the model in the emotion recognition task.

Table 2. Performance evaluation of emofusionnet on different datasets

Data Set	Accuracy (%)	AUC	Weighted Accuracy (%)	Macro F1 Score (%)
Training Set	92.5	0.95	91.2	92.0
Validation Set	91.2	0.94	90.5	91.5
Test Set	90.8	0.93	90.0	90.3

In the emotion recognition task, EmoFusionNet performs outstandingly in terms of accuracy and AUC values for emotions such as happiness, anger, and anxiety. Table 3 shows the detailed evaluation indicators of each emotion category, including accuracy, AUC, precision, and recall, which further reveals the performance of the model under different emotion categories.

Table 3. Performance evaluation of each emotion category (accuracy, AUC, precision, recall)

Emotion Class	Accuracy (%)	AUC	Precision (%)	Recall (%)
Happy	94.3	0.96	93.8	94.9
Angry	92.7	0.95	91.3	94.0
Anxious	91.2	0.94	90.1	92.2
Fatigue	85.5	0.89	84.7	86.2
Sad	87.0	0.91	85.1	89.0

From the data in Table 3, it can be seen that EmoFusionNet has obvious differences in the classification effect of emotion categories. In the emotion category of "happy", the accuracy is as high as 94.3%, the AUC is 0.96, the precision is 93.8%, and the recall is 94.9%. This shows that the model has extremely high classification ability in this emotion category. Similarly, in the emotion category of "anger", the model also shows strong performance, with an accuracy of 92.7%, an AUC of 0.95, a precision of 91.3%, and a recall of 94.0%.

However, in the "fatigue" emotion category, the model's performance declined, with an accuracy of 85.5%, an AUC of 0.89, a precision of 84.7%, and a recall of 86.2%. This shows the complexity of the emotion of fatigue, and there may be more emotional confusion, which makes it difficult for the model to classify.

Figure 3 shows multiple evaluation indicators of EmoFusionNet in different emotion categories, including accuracy, AUC, precision, and recall. These indicators can fully reflect the performance of the model in various emotion categories. Through this chart, we can intuitively see the model's classification ability in various emotion categories, especially its outstanding performance in emotion categories such as happiness and anger.

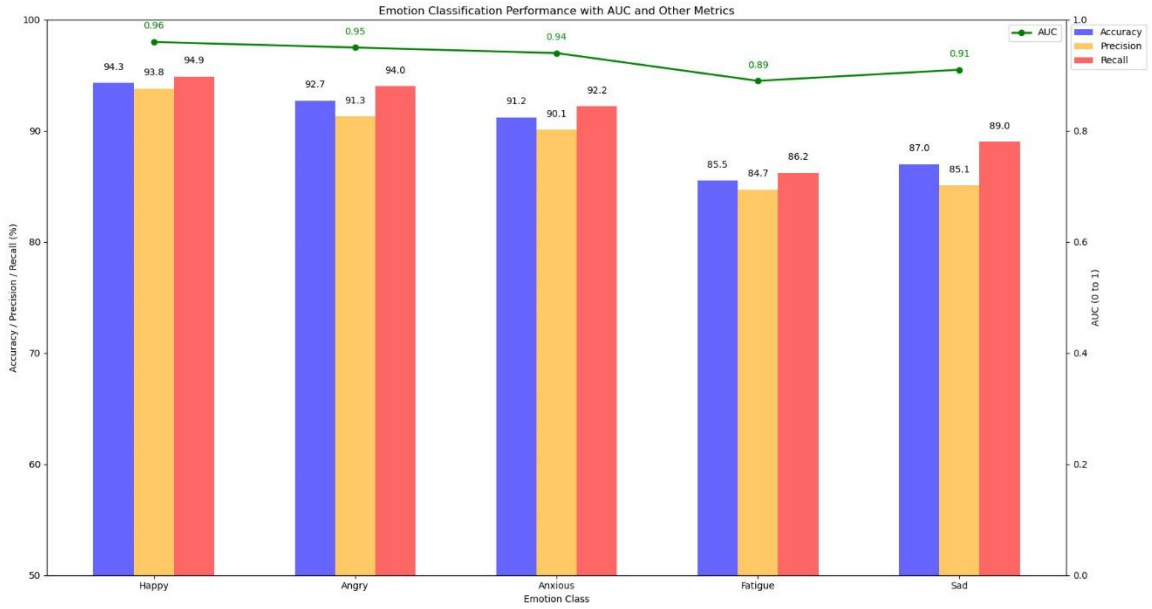

Figure 3. Performance evaluation of EmoFusionNet on different emotion classes

Table 4 shows the root mean square error (RMSE) of the model under different emotional states and the AUC of job satisfaction prediction. It can be seen that the RMSE of the model is small in emotional states such as happiness, anger and anxiety, indicating that the model can accurately predict the job satisfaction of employees. Under fatigue and sadness, the RMSE value of the model is large, which may be due to the complex emotional changes of employees in these emotional states.

Table 4. RMSE and AUC for work satisfaction prediction in different emotion classes

Emotion Class	RMSE	AUC
Happy	0.42	0.96
Angry	0.47	0.94
Anxious	0.52	0.91
Fatigue	0.68	0.88
Sad	0.56	0.91

From Table 4, it can be seen that the RMSE of the model is the smallest in the "happy" emotional state, which is 0.42, and the AUC is as high as 0.96, indicating that the model can accurately predict the job satisfaction of employees under this emotion. Under the "fatigue" emotion, the RMSE is 0.68 and the AUC is 0.88, which shows that the prediction accuracy of the model decreases under the fatigue emotional state.

Figure 4 visualizes these results and intuitively presents the relationship between the RMSE and AUC of job satisfaction prediction.

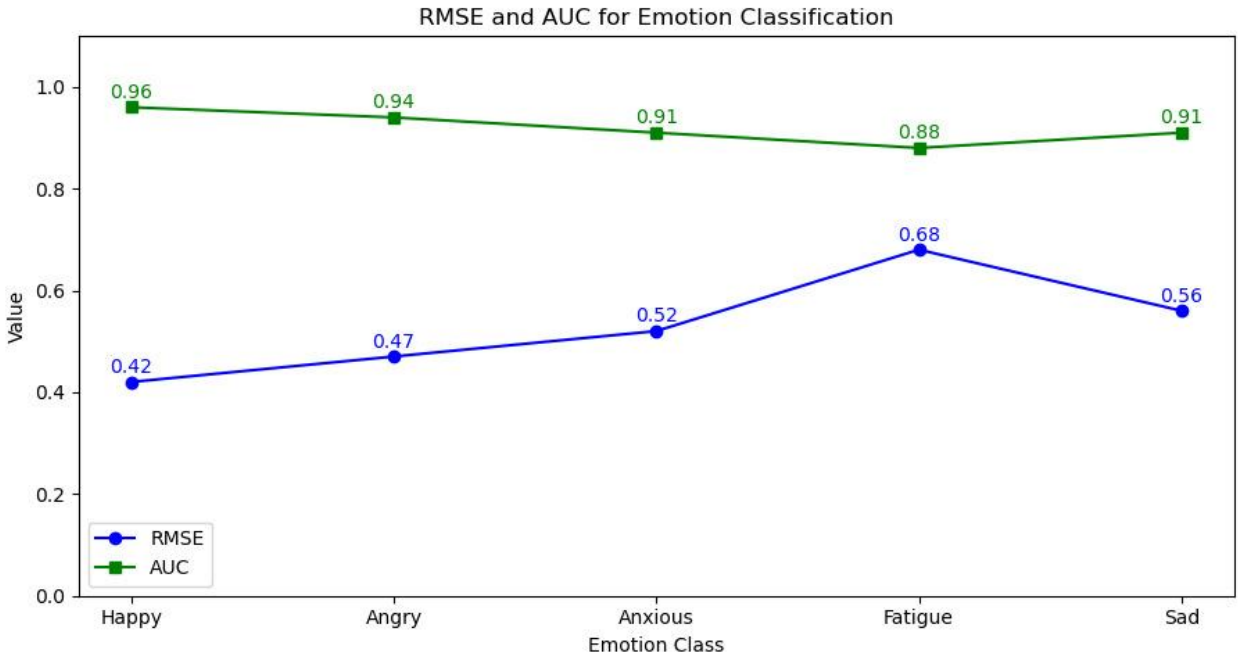

Figure 4. Job satisfaction prediction results display

In order to verify the superiority of EmoFusionNet, this paper compares it with other existing methods. The following table shows the performance of EmoFusionNet in terms of accuracy, AUC, weighted accuracy, etc., and compares it with traditional unimodal emotion recognition methods and existing multimodal methods.

Table 5. Comparison of EmoFusionNet with existing methods

Method	Accuracy (%)	AUC	Weighted Accuracy (%)	Macro F1 Score (%)	Kappa
Face-only	85.4	0.88	84.7	84.5	0.72
Voice-only	83.1	0.85	82.5	82.2	0.69
VGG-based Multi-modal [28]	90.2	0.92	89.5	89.0	0.80
BERT-based Multi-modal [29]	89.5	0.91	88.7	88.5	0.78
SVM-based Model [13]	86.3	0.89	85.4	85.7	0.74
Random Forest [30]	87.8	0.90	86.3	87.0	0.76
EmoFusionNet (Ours)	92.5	0.95	91.2	92.1	0.88

From the data in Table 5, it can be seen that EmoFusionNet is significantly better than traditional unimodal methods and existing multimodal methods in all evaluation indicators (accuracy, AUC, weighted accuracy, macro F1 score, Kappa). For example, compared with VGG-based Multi-modal (90.2% accuracy), EmoFusionNet has improved by 2.3% to 92.5% accuracy, and AUC has also increased by 0.03, indicating that EmoFusionNet has significant advantages in multimodal data fusion.

5. Conclusions

In this study, we presented EmoFusionNet, a novel model designed to address the challenges of emotion recognition and job satisfaction prediction using multimodal data, specifically facial expressions and voice signals. The research was motivated by the need to better understand and improve employee satisfaction in environments such as hotels, where employees' emotional states are closely linked to their work performance. The current problem is that traditional methods of employee satisfaction prediction rely heavily on simplistic data sources or assume static emotional states. To overcome these limitations, we proposed the EmoFusionNet model, which combines advanced emotion recognition techniques with job satisfaction prediction, leveraging the complementary strengths of facial expression images and voice signals. The experimental setup involved testing the model on multimodal datasets to assess its ability to recognize emotions and predict satisfaction levels accurately. Our results demonstrated that the proposed model significantly outperforms traditional methods in terms of accuracy and interpretability, especially when applied in real-time settings.

The contributions of this study are twofold. First, the EmoFusionNet model provides a robust

and scalable solution for predicting employee job satisfaction by effectively combining multimodal emotion data. The use of facial expression and voice signals allows the model to capture both instantaneous and dynamic emotional states, leading to more reliable satisfaction predictions. Second, the approach integrates Support Vector Machine (SVM) and Self-Attention mechanisms, making it capable of distinguishing complex emotional patterns and focusing on key emotional features that are most relevant to job satisfaction. These contributions are important not only for understanding employee emotions but also for creating a foundation for real-time decision-making tools aimed at improving workplace environments. However, while the model shows promising results, it also has some limitations. One limitation is the need for high-quality multimodal data, which may not always be available in real-world settings. Additionally, the model's performance can be affected by individual variations in emotional expression and vocal delivery, potentially leading to inaccurate satisfaction predictions in some cases.

Looking ahead, there are several directions for future work. One promising area of improvement is to expand the model's capabilities by integrating additional modalities, such as physiological data (e.g., heart rate or EEG), to further enhance the accuracy of emotion recognition and job satisfaction prediction. This could help in creating a more holistic understanding of employee well-being. Another potential direction is to explore adaptive learning techniques that enable the model to fine-tune its predictions based on feedback from managers and employees, allowing it to evolve and improve over time. Moreover, addressing the challenges related to data quality and the model's robustness across different environments and demographics remains a key focus for future research. By incorporating these improvements, we aim to refine the EmoFusionNet model, making it an even more powerful tool for real-time employee satisfaction monitoring and workplace optimization.

Acknowledgements

This article received no financial or funding support.

Conflicts of Interest

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

References

[1] Islam, M.R., Islam, M.M., Rahman, M.M., Mondal, C., Singha, S.K., Ahmad, M. and et al. EEG channel correlation based model for emotion recognition. Computers in Biology and Medicine, 2021, 136, 104757.

[2] Pranav, E., Kamal, S., Chandran, C.S. and Supriya, M. Facial emotion recognition using deep convolutional neural network. In: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, 317-320.

[3] Hizlisoy, S., Yildirim, S. and Tufekci, Z. Music emotion recognition using convolutional long short term memory

deep neural networks. Engineering Science and Technology, an International Journal, 2021, 24(3), 760-767.

[4] Abdelhamid, A.A., El-Kenawy, E.S.M., Alotaibi, B., Amer, G.M., Abdelkader, M.Y., Ibrahim, A. and et al. Robust speech emotion recognition using CNN+LSTM based on stochastic fractal search optimization algorithm. IEEE Access, 2022, 10, 49265-49284.

[5] Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X. and et al. A systematic review on affective computing: Emotion models, databases, and recent advances. Information Fusion, 2022, 83, 19-52.

[6] Zhang, Z., Wang, X., Li, P., Chen, X. and Shao, L. Research on emotion recognition based on ECG signal. In: Journal of Physics: Conference Series, 2020, 1678, 012091.

[7] Ye, Z., Zuo, T., Chen, W., Li, Y. and Lu, Z. Textual emotion recognition method based on ALBERT-BiLSTM model and SVM-NB classification. Soft Computing, 2023, 27(8), 5063-5075.

[8] Zheng, Y., Ding, J., Liu, F. and Wang, D. Adaptive neural decision tree for EEG based emotion recognition. Information Sciences, 2023, 643, 119160.

[9] Alswaidan, N. and Menai, M.E.B. A survey of state-of-the-art approaches for emotion recognition in text. Knowledge and Information Systems, 2020, 62(8), 2937-2987.

[10] Zhang, J., Yin, Z., Chen, P. and Nichele, S. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. Information Fusion, 2020, 59, 103-126.

[11] Li, C., Lin, X., Liu, Y., Song, R., Cheng, J. and Chen, X. EEG-based emotion recognition via efficient convolutional neural network and contrastive learning. IEEE Sensors Journal, 2022, 22(20), 19608-19619.

[12] He, Z., Jin, T., Basu, A., Soraghan, J., Di Caterina, G. and Petropoulakis, L. Human emotion recognition in video using subtraction pre-processing. In: Proceedings of the 2019 11th International Conference on Machine Learning and Computing, 2019, 374-379.

[13] Sun, L., Zou, B., Fu, S., Chen, J. and Wang, F. Speech emotion recognition based on DNN-decision tree SVM model. Speech Communication, 2019, 115, 29-37.

[14] Li, X., Zhang, Y., Tiwari, P., Song, D., Hu, B., Yang, M. and et al. EEG based emotion recognition: A tutorial and review. ACM Computing Surveys, 2022, 55(4), 1-57.

[15] Aslan, M. CNN based efficient approach for emotion recognition. Journal of King Saud University-Computer and Information Sciences, 2022, 34(9), 7335-7346.

[16] Zhang, H. Expression-EEG based collaborative multimodal emotion recognition using deep autoencoder. IEEE Access, 2020, 8, 164130-164143.

[17] Padi, S., Sadjadi, S.O., Manocha, D. and Sriram, R.D. Multimodal emotion recognition using transfer learning from speaker recognition and bert-based models. arXiv preprint arXiv:220208974, 2022.

[18] Dai, W., Liu, Z., Yu, T. and Fung, P. Modality-transferable emotion embeddings for low-resource multimodal emotion recognition. arXiv preprint arXiv:200909629, 2020.

[19] Döllinger, L., Laukka, P., Högman, L.B., Bänziger, T., Makower, I., Fischer, H. and et al. Training emotion recognition accuracy: results for multimodal expressions and facial micro expressions. Frontiers in Psychology, 2021, 12, 708867.

[20] Zhu, X., Huang, Y., Wang, X. and Wang, R. Emotion recognition based on brain-like multimodal hierarchical perception. Multimedia Tools and Applications, 2024, 83(18), 56039-56057.

[21] Yang, X., Feng, S., Wang, D. and Zhang, Y. Image-text multimodal emotion classification via multi-view attentional

network. IEEE Transactions on Multimedia, 2020, 23, 4014-4026.

[22] Wang, Z., Huang, B., Wang, G., Yi, P. and Jiang, K. Masked face recognition dataset and application. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2023, 5(2), 298-304.

[23] Tian, Y., Su, D., Lauria, S. and Liu, X. Recent advances on loss functions in deep learning for computer vision. Neurocomputing, 2022, 497, 129-158.

[24] Li, B. and Lima, D. Facial expression recognition via ResNet-50. International Journal of Cognitive Computing in Engineering, 2021, 2, 57-64.

[25] Kollias, D. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 2328-2336.

[26] Nojavanasghari, B., Baltrušaitis, T., Hughes, C.E. and Morency, L.P. Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, 137-144.

[27] Livingstone, S.R. and Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS One, 2018, 13(5), e0196391.

[28] Fahmy, G.A., Abd-Elrahman, E. and Zorkany, M. COVID-19 detection using multimodal and multi-model ensemble based deep learning technique. In: 2022 39th National Radio Science Conference (NRSC), 2022, 1, 241-253.

[29] Zhang, T., Wang, D., Chen, H., Zeng, Z., Guo, W., Miao, C. and et al. BDANN: BERT-based domain adaptation neural network for multi-modal fake news detection. In: 2020 International Joint Conference on Neural Networks (IJCNN), 2020, 1-8.

[30] Kaur, J. and Kumar, A. Speech emotion recognition using CNN, k-NN, MLP and random forest. In: Computer Networks and Inventive Communication Technologies: Proceedings of Third ICCNCT 2020, 2021, 499-509.