**Journal of Information Systems Engineering and Management**
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

# Emotional Analysis using Spiking Neural Networks

[1]Kotamareddi Abigna, [2]Manju.G, [3]Srisakthi Saravanan
[1]*Computer Science and Engineering department Vellore Institute of Technology, Chennai, India*
abigna.2003@gmail.com
[2]*School of Computer Science and Engineering Vellore Institute of Technology Chennai, India*
manju.g@vit.ac.in
[3]*School of Computer Science and Engineering Vellore Institute of TechnologyChennai, India*
srisakthi.saravanan@vit.ac.in

**ARTICLE INFO**

Received: 29 Dec 2024

Revised: 12 Feb 2025

Accepted: 27 Feb 2025

**ABSTRACT**

With Artificial Intelligence gaining popularity in every field. Emotions is the one thing that Artificial Intelligence or AI cannot replicate yet. One of the key reasons being that emotions are often complex and hard to understand even by humans who used to them. Not only are these emotions complex and hard to understand, humans often express their emotions in different ways. Some don't even express it at all. While some, express it using big motions and exaggerated facial expressions. There're also cases where it is expressed in the voice via small changes in the pitch and frequency of the words as well as the intonation. All of these different ways of expression as well as the complications of these emotions make it hard for it understood much less replicated. Emotional analysis is method to understanding emotions that are expressed by humans in various different forms. Emotional analysis is a popular problem statement, and is constantly worked on with the advancing technology. One such new technology is Spiking Neural Networks, a newer model of Neural Networks that is based on the biological spiking of neurons to pass information in the brain. In this paper, we propose the method of using SNN's on multi-modal data for the purpose of emotional analysis. The multi-modal data that we used encompasses both physiological and physical signs of emotion. We have also tested uni-modals, bi-modals and multi modals of the same data.

**Keywords:** spiking neural networks, emotional analysis, neural networks, emotional analysis with neural networks, and emotional analysis with spiking neural networks

## I. INTRODUCTION

Emotional analysis is now an interesting area of research because technology now intersects and interacts often with human behaviour, cognition. Often times even going as far to impact the emotional well-being of humans. Within the past several years, widespread use of multimedia content, social media engagement, and wearable sensors has generated humongous amounts of data regarding the state of a person in multifaceted and composite forms. Conventional approaches in affective computing, based typically on static analysis of data and traditional neural networks, have gone a long way in deciphering these emotional cues. But they tend to fall short in simulating the dynamic, time-variant character of human emotions and are not as energy-efficient as is required by the task of intervention in sensitive cases.

Spiking Neural Networks (SNNs) are a promising contender, providing a paradigm shift in the way computational models can simulate the complex dynamics of biological neural systems. SNNs differ from conventional deep learning models that handle information statically or rate-coded, as they simulate the temporal dynamics of actual neurons by representing information in the exact timing of electrical spikes. This temporal coding is a more accurate representation of time-varying emotional cues, allowing for a more realistic emulation of the way human brains sense, process, and react to emotional stimuli. The built-in energy efficiency of SNNs, which stems from their event-driven processing, also makes them appealing for real-time processing in environments with limited computational resources.

The incorporation of SNNs into emotional analysis not only closes the gap between neuroscience and artificial intelligence but also remedies some of the issues raised by classical models. For example, whereas conventional deep

learning techniques tend to demand tremendous amounts of data and immense computational resources, SNNs have the potential to decrease these demands in terms of sparse and asynchronous neuron communication. This transition is especially applicable in domains like wearable health monitoring, robotics, and interactive systems, where efficient and timely emotion recognition is critical.

This paper is a thorough overview of SNN-based methods to emotional analysis. We start with the theoretical framework of spiking neural networks and their neurobiological basis, emphasizing how such models can simulate the temporal and dynamic nature of emotional expression. Finally, we explore the methodology used in our research, describing the experimental setup, datasets, and performance measures utilized to compare the SNN models with traditional deep learning models. We pay special care to the preprocessing methods required to prepare noisy and high-dimensional emotional data and to the techniques used to maximize the network's spiking patterns towards improved accuracy.

In addition, the study explores the practical limitations of applying SNNs, such as challenges in training the networks and making trade-offs between biological plausibility and computational efficiency. Comparing the performance of SNNs with state-of-the-art models on varied emotional datasets, this work hopes to determine when SNNs have unique benefits and suggest remedies for their shortfalls.

In this work, we attempt to contribute to the changing dynamics of affective computing by illustrating how models inspired from biology can result in more accurate, responsive, and resource-saving emotion recognition systems. The results shown in this paper not only emphasize SNNs' potential to revolutionize emotional analysis but also give directions to future work that combines lessons from neuroscience with sophisticated machine learning methodologies.

## BACKGROUND AND MOTIVATION

The fast pace of technological development and data availability has spurred novel strategies for understanding human emotions. As emotional analysis plays a more central role in fields such as human–computer interaction, psychology, and artificial intelligence, novel computational models are needed to model the richness and volatility of human affect. This section discusses the history of emotional analysis, the genesis of Spiking Neural Networks (SNNs) as a biologically inspired approach, and the motivations for ongoing research in this area.

### 2.1 Emotional Analysis: A Growing Field

Human emotions are complex, dynamic, and context-specific. Historically, emotional analysis has been approached using techniques like sentiment analysis, facial recognition, and physiological signal monitoring to predict affective states. With the abundance of multimedia data—from social media posts to video recordings—researchers began to build models that not only recognize but also predict patterns of emotions over time. But traditional models usually fail with:

Temporal Dynamics: Emotions change over time, yet most models analyze data in a static way.

Contextual Nuances:The nuanced differences in human expression need to be captured by advanced representations beyond mere classification.

Data Heterogeneity: The diverse sources and forms of emotional data (e.g., text, audio, video) make it difficult to standardize analysis techniques.

These issues point toward the requirement of methods that are able to process temporal information and the subtlety of affect expression in a native manner.

### 2.2 Shortcomings of Standard Methods

Standard neural network designs—deep learning models, for instance—have achieved profound breakthroughs in pattern classification and categorization tasks. However, their design often follows rate-coded representations and rigid processing paradigms. This results in the following shortcomings when applying them to emotional analysis:

Temporal Non-Precision: Most conventional models fail to capture the delicate, time-varying patterns in emotional cues.

**Journal of Information Systems Engineering and Management**
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Computational Costliness: Deep neural models are usually computationally intensive, and this might be prohibitive for real-time and embedded systems.

Biological Realism:Conventional models do not mimic biological spiking temporal behavior of neurons and thus may not be able to effectively replicate human emotional processing.

These limitations spur the investigation of alternative models that better match how biological systems operate.

2.3 Spiking Neural Networks: A Bio-Inspired Approach

Spiking Neural Networks (SNNs) are a dramatic departure from traditional neural network structures. Modeled after how neurons talk to each other in the brain, SNNs represent information in the form of discrete spikes instead of continuous activations. Salient aspects of SNNs are:

Temporal Coding: Through the use of the exact timing of spikes, SNNs encode temporal dynamics that are essential to grasp changing emotional states.

Event-Driven Processing: SNNs process on demand, only firing when important events take place. This results in higher energy efficiency—a key benefit for real-time systems.

Biological Realism:The spiking nature of SNNs resembles that of real neural circuits and can possibly shed light on the inherent processes of human emotion.

The use of SNNs in emotional analysis can not only guarantee greater accuracy in the recording of temporal patterns but also a more efficient and physiology-oriented computational platform.

2.4 Motivation for Integrating SNNs in Emotional Analysis

The inclusion of SNNs in emotional analysis is motivated by theoretical and practical reasons. The motivation for the research can be described as follows:

Temporal Nuances: Emotions are time-evolving, and SNNs, due to their ability to temporally code, are particularly apt at modeling these dynamics. This would result in more precise and context-sensitive emotion recognition systems.

Efficiency of Resources:Since SNNs operate information processing in an event-based mode, they potentially provide drastic computational overhead savings. This efficiency can be of immense value in wearable devices, robots, and other applications where power usage is a factor.

Bridging Biological and Computational Models: By matching the computational models to neurobiological processes, SNNs bring a framework not only functionally efficient but conceptually consistent with the way in which humans learn and process emotionally charged stimuli. This can inform greater understanding into artificial as well as biological intelligence.

Overcoming Existing Constraints:Standard models typically need enormous sets of annotated data and are challenged by noise and variability in emotional expression. The adaptive, asynchronous architecture of SNNs may provide resilience against these imperfections, which would open the doors to more fault-tolerant emotion recognition systems.

**C. Significance of the work**

The value of this work is that it has the potential to push affective computing forward by combining Spiking Neural Networks' unique strengths with emotion analysis. The following highlights its value:

Improved Temporal Dynamics and Biological Plausibility

SNNs naturally represent information in the form of the exact timing of neural spikes, very closely modeling the biology of neurons. This ability allows temporal dynamics of emotional expression to be modeled with orders of magnitude higher fidelity than classical DNNs, which frequently depend on rate-coded or static representations. With the ability to capture the fine-grained time-evolution of emotional states, SNN-based systems can support more accurate and context-sensitive analysis—a requirement imperative for tasks such as real-time sentiment tracking and adaptive human-machine interaction.

Energy Efficiency and Real-Time Performance

**Journal of Information Systems Engineering and Management**
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

One of the major benefits of SNNs is their event-driven processing, which leads to substantial energy savings. This energy efficiency is especially critical for embedded systems, wearable devices, and mobile platforms, where computational resources and battery life are limited. The reduced power demands, along with the possibility of real-time processing, make SNNs an appealing choice for implementing emotion recognition systems in resource-limited environments, thus extending the range of their practical applications.

Multimodal Integration and Robustness

Emotion identification in natural applications frequently involves blending information from disparate sources, i.e., EEG signals, voice, and face. The empirical studies reviewed establish that SNNs can appropriately combine multimodal information, promoting more resilient and stable performance under conditions of noise and data deterioration. Such multimodality is not only more accurate in classifying but is also more tolerant to failure by guaranteeing effective performance across the variety of operating conditions and population of users.

Bridging the Gap Between Neuroscience and Machine Learning

Through its alignment of computational models with neurobiological processes, this work advances the understanding of how human emotions may be computationally modeled. The biologically inspired SNN design acts as a bridge between neuroscience and machine learning, resulting in interdisciplinary insights potentially driving future innovations in each area. Such integration is poised to create next-generation systems not only efficient and accurate but also offer a glimpse into the hidden mechanisms of human emotional processing.

Implications for Future Research and Applications

Successful integration of SNNs in emotion analysis presents several potential avenues for future research. Advancements in training methods, design of hybrid approaches fusing SNNs and traditional deep learning, and extension into multimodal fusion of data will increase performance and practical applicability of affect recognition systems. In addition, the low power consumption characteristics of SNNs make them an ideal choice for real-time solutions for fields such as robotics, medicine, and interactive gaming, where it becomes progressively essential to sense and interpret human emotion.

In total, the importance of this work is that it has the capability to breach the limitations that exist within conventional emotion recognition approaches while providing a route toward more efficient, more accurate, and biologically informed computational models. This development will have long-term implications in a wide range of fields, and it will be pushing innovation in theoretical investigation as well as in real-world applications in affective computing.

## RELATED WORKS

Current research in emotion analysis has increasingly relied on Spiking Neural Networks (SNNs) due to their ability to encode temporal dynamics, simulate biological neural activity, and offer energy-efficient computation. Several studies have explored SNN applications across multiple modalities—EEG, speech, facial expressions, and multimodal fusion—to address limitations in traditional deep learning paradigms.

SNNs in Theoretical and Practical Perspectives

Yamazaki et al. (2022) [1] provide a comprehensive overview of SNN architectures, comparing their energy efficiency and temporal processing capabilities to traditional deep neural networks (DNNs). Tavanaei et al. (2019)[2] discuss challenges in combining deep learning techniques with SNNs, particularly the non-differentiability issue and the need for surrogate gradient methods. These foundational works emphasize the biological plausibility and computational benefits of SNNs for emotion analysis.

EEG-Based Emotion Recognition

Several researchers have leveraged SNNs for EEG-based emotion recognition. Luo et al. (2020)[5] compared signal processing methods such as discrete wavelet transform, variance, and fast Fourier transform, demonstrating that SNNs outperform conventional classifiers in capturing spatiotemporal EEG patterns. Alzhrani et al. (2021)[6] employed the NeuCube framework—a brain-inspired SNN structure—to classify spatiotemporal EEG data with high

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

accuracy. Li et al. (2023)[13] introduced a Fractal-SNN paradigm that exploits multi-scale temporal-spectral-spatial characteristics, further enhancing EEG data utilization for emotion recognition.

Speech Emotion Recognition

In speech emotion recognition, Mansouri-Benssassi and Ye (2021) [3] proposed an SNN model incorporating early cross-modal enhancement, inspired by the auditory processing of the brain, to improve dynamic speech signal modeling. Jain and Shukla (2022) [12] integrated deep belief networks for feature learning with an SNN-based decision-making architecture, significantly improving performance over traditional approaches. These studies illustrate the sensitivity of SNNs in capturing transient emotional cues in speech signals.

Facial Expression Recognition and Dynamic Vision

SNN-based methods have also contributed to facial expression recognition. Fu et al. (2021)[15] developed a cortex-like SNN model that mimics hierarchical visual cortex organization, effectively recognizing complex facial expressions. Barchid et al. (2023)[14] introduced "Spiking-FER," an event-based SNN model optimized for data from dynamic vision sensors, achieving competitive accuracy with lower energy consumption compared to artificial neural networks (ANNs). These findings underscore the suitability of SNNs for real-time, energy-efficient facial expression analysis.

Multimodal and Multisensory Integration

Given the inherently multimodal nature of human emotions, researchers have explored fusion approaches using SNNs. Tan et al. (2021)[10] employed the NeuCube framework to integrate facial expressions and physiological signals, yielding robust emotion classification. Mansouri-Benssassi and Ye (2021)[7] demonstrated that SNN-based models effectively fuse auditory and visual data to enhance emotion recognition performance. These multimodal approaches not only improve classification accuracy but also enhance resilience to noise and data degradation.

Sentiment Analysis and Other Modalities

Beyond emotion recognition, SNNs have been applied to sentiment analysis. Chunduri and Perera (2023)[11] introduced a neuromorphic sentiment analysis model running on SpiNNaker hardware, achieving high accuracy with low energy consumption. This research expands the applicability of SNNs into natural language processing (NLP), further broadening their role in affective computing applications.

SNNs have demonstrated significant potential across various domains of emotion analysis, including EEG, speech, facial expressions, and multimodal fusion. Their ability to efficiently process temporal and spatiotemporal data, coupled with lower energy consumption, positions them as a powerful alternative to traditional deep learning models in affective computing. Future research should focus on improving training techniques, enhancing multimodal integration, and optimizing neuromorphic hardware implementations to fully harness the capabilities of SNNs.

## II. PROPOSED METHODOLOGY

The suggested approach to multimodal emotion recognition with Spiking Neural Networks (SNNs) combines audio, video, and EEG signals to classify emotions into three general categories: negative, neutral, and positive. The datasets employed are RAVDESS for speech and video data, MELD for conversational sentiment analysis, and an EEG brainwave dataset for physiological signals. The methodology starts with preprocessing, where emotional labels from various datasets are normalized to ensure consistency. In the audio stream, Mel-frequency cepstral coefficients (MFCCs) are obtained to represent the spectral features of speech signals. In the video stream, a pre-trained ResNet-18 model is utilized to extract deep feature representations from facial expressions. At the same time, EEG signals are treated by averaging and normalizing the values into fixed-length vectors for uniformity among recordings. This preprocessing guarantees that the three modalities are properly organized prior to being input into the learning pipeline.

This is necessary to avoid class imbalance, which might otherwise skew the model in favor of more common emotional classes. A PyTorch dataset class is created to effectively manage multimodal inputs so that audio, video, and EEG features are easily integrated. The dataset is then divided into training and test sets, and a PyTorch DataLoader is utilized to load data batches during training and testing. This organized data pipeline is crucial in

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

optimizing computational efficiency while preserving the integrity of multimodal information. At the heart of the model lies a Recurrent Spiking Neural Network (RecurrentSNN), which is trained on input data over multiple time steps (T).

The network is comprised of fully connected layers interspersed with Leaky Integrate-and-Fire (LIF) recurrent cells, which replicate biologically inspired neuronal dynamics. The LIF model, through the use of Norse (a PyTorch-based library for SNNs), adds dynamic spiking activity that enables the model to extract temporal dependencies from the multimodal data. In addition, a different SNN architecture is investigated in which sequential LIF layers are used instead of the recurrent architecture to compare performance between various network topologies. Through the utilization of spiking neuron dynamics, the model tries to extract static and dynamic emotional cues in the multimodal inputs. Cross-entropy loss is used as the objective function during training, while the Adam optimizer is used to update network parameters for 50 epochs.

The loop of training consists of forward pass in the SNN, calculating loss as a function of predicted versus actual emotional labels, and backpropagation of gradients to update weights. Model performance while testing is measured using accuracy measurements, classification reports, and confusion matrices. Results analysis is done in great detail through visual methods such as heatmaps, giving insights into how well the network can differentiate between various emotional classes. This approach takes advantage of the specific strengths of SNNs, including their energy efficiency and temporal processing capabilities, to provide a new solution for multimodal emotion recognition.

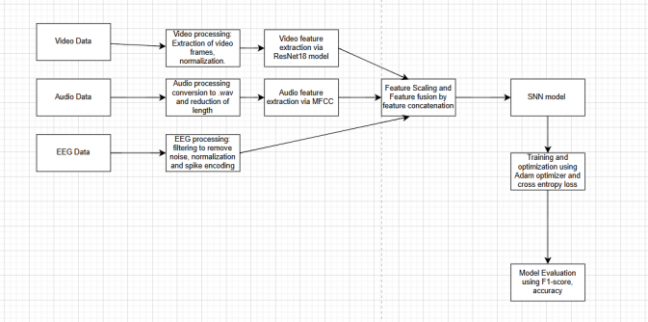



**Fig-1 Proposed Architecture**

## III. METHODOLOGIES

The paper presents a multimodal emotion recognition system based on EEG, video, and audio modalities. The system integrates a Spiking Neural Network (SNN) to perform resilient feature fusion and classification. Data acquisition, preprocessing, feature extraction, model design, training, and evaluation are the steps employed in the methodology. The aim is to maximize emotion recognition accuracy using SNNs' efficiency in processing temporal dependencies.

I. Data Acquisition

We make use of three publicly released datasets, each recording various modalities of emotion expression:

RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song): Holds emotional speech and song recordings of 24 actors, annotated with diverse emotions.

MELD (Multimodal EmotionLines Dataset): A sentiment analysis dataset, comprising video and text transcripts derived from TV show dialogues.

EEG Brainwave Dataset: Contains EEG signals captured from subjects undergoing varying emotional states, providing physiological understanding of emotion response.

II. Data Preprocessing

A. Label Mapping

Emotion labels within datasets are mapped into three general categories to make the fusion of features and classifying of the emotions easier:

**Journal of Information Systems Engineering and Management**
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

Negative (0): Comprises anger, disgust, sadness, and fear.

Neutral (1): Comprises calm and neutral states.

Positive (2): Comprises happiness, joy, and surprise.

Due to constraints of the publicly available datasets, label mapping was required.

B. Feature Extraction

We extract meaningful features from every modality to form a common multimodal representation:

Audio Processing: Mel-Frequency Cepstral Coefficients (MFCCs) are used for audio processing, converting raw waveform data into a concise, discriminative feature set.

Video Processing: A ResNet-18 model is used, pre-trained, to extract deep vision features, representing spatial patterns that are important for facial expressions.

EEG Processing: EEG signals are processed and transformed into fixed-size vectors (128 size) by averaging the signal over time, making them consistent.

C. Dataset Balancing

To handle class imbalance, we use oversampling techniques that duplicate minority samples, achieving a balanced dataset that enhances model generalization.

IV. Model Architecture

A. Multimodal Fusion

Features extracted from audio (40-dimensional MFCCs), video (512-dimensional ResNet features), and EEG (128-dimensional vectors) are concatenated, creating a combined 680-dimensional feature vector. The fused representation represents varied emotional cues across modalities.

B. Spiking Neural Network (SNN) Model

The Spiking Neural Network (SNN) is intended to handle multimodal inputs through the mimicking of biological neurons' spiking activity. SNNs have energy efficiency and temporal data processing benefits which are suitable for the task of emotion recognition. We investigate two SNN architectures:

1. Recurrent SNN

This architecture uses a Leaky Integrate-and-Fire (LIF) Recurrent Cell that stores information across time steps, mimicking biological neural dynamics. It learns temporal dependencies through integrating spikes over ime steps, making effective temporal feature learning possible. The recurrent nature of this SNN enables sequential pattern capture in EEG signals and audio waveforms.

2. Feedforward SNN

This model has two layers of LIF neurons, which bring sparsity to network activation, reducing energy consumption. SNNs are different from conventional deep networks in that they are based on spike timing instead of continuous activations, enhancing computational efficiency. The feedforward structure is especially beneficial for combining static video features with dynamic EEG and audio representations.

3. Neuron Dynamics and Encoding

To handle inputs in an event-based fashion, the rate coding method is employed, with the highest spike rates indicating stronger activations. This enables the SNN to represent variation in intensity in emotions across modalities. Spike-time-dependent plasticity (STDP) mechanisms are also under consideration for possible adaptive learning.

V. Training and Optimization

Training is done with:

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

Loss Function as Cross-Entropy Loss which is appropriate for multi-modal classification. We have used the Adam optimizer with learning rate 0.001, providing stable convergence. We have chosen a Batch Size of 16, maximizing memory efficiency and gradient updates.

During training, gradient-based backpropagation through time (BPTT) with surrogate gradients is used, enabling efficient weight updates in spite of SNNs' discrete spiking nature. The surrogate gradient method assists in overcoming the non-differentiability of spike activation functions.

VI. Evaluation

The trained model is assessed using accuracy, classification report and confusion matrix. The model is also compared to two different ANN models as well as single modality SNNs for all three modalities and video + audio modality SNN.

Results indicate how the various modalities contribute to general classification accuracy with EEG signals acting as a prime factor in segregating fine emotional states.

## IV. RESULTS

| | EEG | RAVDESS AUDIO | VIDEO | AUDIO+VIDEO | AUDIO+VIDEO+EEG |
|---|---|---|---|---|---|
| SNN | 60% | 85.07% | 50% | 81% | 90% |

| | AUDIO+VIDEO+EEG |
|---|---|
| SNN | 90% |
| ANN | 68.73% |
| Recurrent ANN | 83.39% |

| | MELD AUDIO | MELD AUDIO+VIDEO+EEG |
|---|---|---|
| SNN | 47.75% | 48.93% |

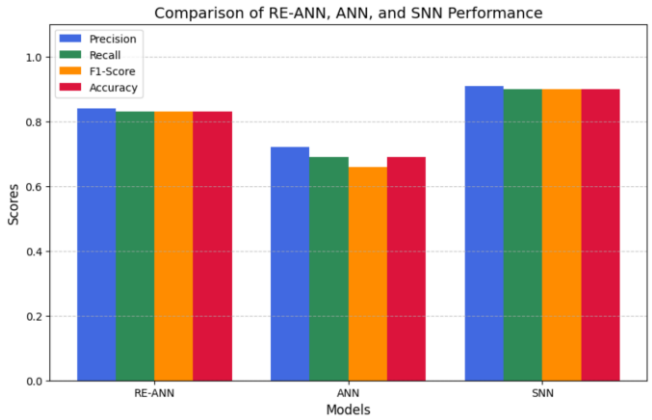
| Model | PRECISION | RECALL | F1-SCORE | ACCURACY |
|---|---|---|---|---|
| RANN | 0.84 | 0.83 | 0.83 | 0.83 |
| ANN | 0.72 | 0.69 | 0.66 | 0.69 |
| SNN | 0.91 | 0.90 | 0.90 | 0.90 |

| Sentiment | Model | Precision | Recall | F1-score | Support |
|---|---|---|---|---|---|
| Negative | RANN | 0.85 | 0.68 | 0.76 | 319 |
| | ANN | 0.83 | 0.31 | 0.45 | 319 |
| | SNN | 0.97 | 0.80 | 0.87 | 319 |
| Neutral | RANN | 0.81 | 0.96 | 0.88 | 319 |

**Journal of Information Systems Engineering and Management**
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

| | | | | | |
|---|---|---|---|---|---|
| | ANN | 0.71 | 0.88 | 0.79 | 319 |
| | SNN | 0.87 | 0.96 | 0.91 | 319 |
| Positive | RANN | 0.84 | 0.86 | 0.85 | 319 |
| | ANN | 0.63 | 0.87 | 0.73 | 319 |
| | SNN | 0.89 | 0.96 | 0.92 | 319 |




The comparison of models and modalities for emotion recognition emphasizes the strengths of various neural network architectures in handling EEG, audio, and video data. Emotion recognition is a difficult task that can be greatly enhanced by multimodal fusion since single modalities usually only capture partial information.

The experiments show that Spiking Neural Networks (SNNs), Artificial Neural Networks (ANNs), and Recurrent Artificial Neural Networks (RANNs) perform differently when implemented on various datasets and modality pairs, highlighting the need for appropriate choice of architecture and feature fusion approach. In performance testing of SNNs on separate modalities, there is evidently a difference in accuracy with respect to the nature of data to be handled.

EEG-based recognition of emotions using SNN has an accuracy rate of 60%, which is moderate but inferior to the rate of audio-based recognition, with a rate of 85.07%. The worst among the three is video, with a mere 50% accuracy rate, indicating facial expressions might be inadequate for a strong emotion classification. This may be due to factors like differences in lighting, occlusions, and the fineness of facial expressions for some emotions. The success of audio-based recognition implies that speech characteristics like tone, pitch, and frequency changes are imbued with dense emotional content. But the actual benefit occurs when modalities are integrated. The combination of video and audio enhances accuracy to 81%, meaning that video on its own is not very reliable, but in combination with speech features, it adds complementary information. The most remarkable enhancement comes when EEG is incorporated into the combination, giving a maximum accuracy of 90%. This validates that brain activity information, when integrated with the conventional sensory inputs, improves the system's capability to detect profound emotional states that might not be verbally or facially expressed. Another comparison among various neural network architectures for multimodal emotion recognition indicates that SNNs perform better than both ANNs and RANNs when all three modalities of audio, video, and EEG are integrated. SNNs have a 90% accuracy, and RANNs trail closely with 83.39%, whereas ANNs fall far behind with 68.73%.

SNNs' high performance lies in their ability to process events, which helps them naturally accommodate spatiotemporal patterns in multimodal data. As opposed to regular ANNs that process data within static frames, SNNs utilize biologically inspired mechanisms for processing information dynamically, resulting in improved feature extraction and decision-making. RANNs, intended to pick up temporal dependencies, perform significantly better than ANNs, reaffirming that sequential modeling is essential for emotion detection. Nevertheless, they are still not better than SNNs, suggesting that SNNs' special architecture offers a computational benefit for managing multimodal

Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

fusion. The experiment on the MELD dataset also demonstrates how different dataset properties influence model performance. When used on audio-based recognition in MELD, SNNs reach a mere accuracy of 47.75%, which is well below their result on RAVDESS audio data. Even when the video and EEG are added to it, accuracy increases marginally to 48.93%.

This small gain indicates that either the fusion approach employed was not the best or that the MELD dataset is more challenging in some way, e.g., more variability in speaker expressions, background noise, or conversational complexity.

These findings are diametrically opposite to the high accuracy found on the RAVDESS dataset, suggesting that various datasets need customized processing methods to deliver best results.

The low performance on MELD indicates that multimodal fusion by itself is inadequate if the contributing modalities are not providing useful information or if the dataset is more difficult in nature. In general, this comparison highlights the significance of choosing the appropriate architecture and modality pair for emotion recognition. The results show that multimodal fusion dramatically improves accuracy, and EEG is essential for boosting classification performance when fused with audio and video. SNNs consistently outperform other architectures, making them a promising solution to multimodal emotion recognition because they are efficient at processing spatiotemporal patterns.

Yet dataset attributes are important in deciding model performance, as evident from the disparate outcomes between RAVDESS and MELD.

Optimizing fusion methods, enhancing EEG feature extraction, and incorporating more sophisticated architectures are recommended areas of exploration in subsequent studies to increase robustness with various datasets. The outcome underscores the importance of a versatile solution that adjusts to the specific challenges introduced by various datasets while capitalizing on the strengths of multimodal learning.

## V. CONCLUSION

Multimodal emotion recognition from EEG, audio, and video is an emerging field of research that aims to narrow the gap between human emotional expression and machine interpretation. The comparative analysis of various neural network architectures—specifically Spiking Neural Networks (SNNs), Artificial Neural Networks (ANNs), and Recurrent Artificial Neural Networks (RANNs)—establishes the drastic effect of both architecture selection and modality combination on classification performance. The findings show that although unimodal methods yield significant information on emotion recognition, multimodal integration results in a dramatic improvement in accuracy.

The performance comparison on the RAVDESS dataset identifies the strengths and limitations of various modalities. EEG-based emotion recognition with SNNs had a moderate accuracy of 60%, which is lower than audio-based classification at 85.07%. This implies that EEG alone, as much as it records brain patterns associated with emotion, might take more sophisticated preprocessing and feature extraction methods to reach its full potential. Alternatively, video-based emotion recognition was poorest at 50%, suggesting that facial expressions were not a potent enough indicator on their own to accurately classify emotion. But when the audio and video were fused together, the accuracy was 81%, again emphasizing the value of multimodal fusion. The best accuracy was when EEG was added to the audio-video fusion to 90%, which shows that brain activity data adds the depth and reliability needed for emotion classification models.

An additional comparison of alternative neural network structures on the composite audio, video, and EEG data set also validates that SNNs perform better than ANNs and RANNs. SNNs achieved a remarkable accuracy of 90%, whereas RANNs were next at 83.39%, and ANNs lagged at 68.73%. This gap in performance indicates that SNNs' dynamic processing of spatiotemporal patterns effectively renders them best suited for emotion recognition tasks. In contrast to traditional ANNs based on static frames, SNNs take advantage of biologically motivated event-driven processes for efficient data processing. The performance of RANNs being considerably higher than ANNs also underscores the significance of sequence modeling in emotion recognition, given that emotions tend to change over time instead of being recorded in a single frame or audio segment.

Performance on the MELD dataset is a different story altogether. Applying SNNs to MELD audio data yielded an accuracy of just 47.75% and an incremental rise to 48.93% when adding EEG and video. This dramatic difference from RAVDESS performance indicates the significant influence of dataset features on model behavior. MELD includes conversational data with higher variability in speaker expressions, background noise, and linguistic details, which could need more advanced fusion methods and feature extraction strategies. That multimodal fusion did not produce much improvement in MELD suggests that modalities are not always better together if the data quality or the fusion approach is less than ideal.

In spite of these encouraging results, however, there are also some challenges that remain to be overcome. EEG signals are noisy and user-dependent, which makes generalization across different users challenging. Standardizing the methods of collecting EEG data and enhancing feature extraction algorithms would serve to reduce this limitation. Further, though multimodal fusion has been found to improve classification accuracy, there remains a necessity for more sophisticated fusion methods that weigh the contribution of each modality dynamically instead of equally. Deep learning models that involve attention mechanisms or reinforcement learning-based fusion could further improve the system's flexibility.

Real-time emotion recognition is another major area for potential study in the future. Many existing models, including those covered in this work, are applied directly to pre-recorded datasets with offline processing. For real-world applications like human-computer interaction, mental health monitoring, and affective computing, real-time processing is necessary. Employing lean, mean, SNN models with low latency that can process multimodal data in real-time would be a major breakthrough. Hardware acceleration via neuromorphic computing platforms such as Intel Loihi or SpiNNaker can potentially further improve SNN-based emotion recognition systems for deployment in real-world applications.

In addition, dataset diversity and generalizability are still major issues. The RAVDESS dataset, as useful as it is for controlled experiments, is unlikely to capture the richness of real-world emotions. More spontaneous and diverse emotional expressions in datasets should be incorporated to enhance the robustness of models. Transfer learning methods can also be investigated to transfer models trained on one dataset for use on another to enhance adaptability from one environment to another and from one population to another.

Ethical and privacy issues also have to be dealt with as emotion recognition technology develops. EEG and facial expression information are very personal data, and securing safe storage and processing of data is of paramount importance. Future studies should aim to design privacy-friendly methods, like federated learning, where models can be trained on decentralized devices without exposing raw data. This would enhance both security and scalability, making emotion recognition systems more widely adoptable in real-world scenarios.

Lastly, the fusion of multimodal emotion recognition with other AI-based applications is full of promising possibilities. Merging emotion recognition and natural language processing (NLP) for sentiment analysis may enable more empathetic virtual helpers and chatbots. In the same way, integrating these systems in healthcare may facilitate early diagnosis of mental illnesses like depression and anxiety. Emotion-aware AI may also be used in autonomous systems such as robots and driverless vehicles to enhance human-machine interactions.

In summary, although the results of this work show the strength of multimodal fusion and the dominance of SNNs for emotion recognition, there are still some challenges and areas where more improvement can be done. The future work must include improving EEG processing methods, optimizing multimodal fusion methodologies, maintaining real-time adaptability, and considering ethical aspects. By overcoming these challenges, researchers can lay the foundation for more accurate, more robust, and more ethical emotion recognition systems that are applicable on a large scale across human-computer interaction, healthcare, and beyond.

## REFERENCES

[1] Yamazaki, K., Vo-Ho, V. K., Bulsara, D., & Le, N. (2022). Spiking neural networks and their applications: A review. Brain Sciences, 12(7), 863. https://doi.org/10.3390/brainsci12070863

[2] Tavanaei, A., Ghodrati, M., Kheradpisheh, S. R., Masquelier, T., & Maida, A. (2019). Deep learning in spiking neural networks. Neural Networks, 111, 47–63. https://doi.org/10.1016/j.neunet.2018.12.002

[3] Mansouri-Benssassi, E., & Ye, J. (2021). Speech emotion recognition with early visual cross-modal enhancement using spiking neural networks. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) (pp. 3127–3134).
[4] Yu, X., Wang, F., & Qiao, Z. (2022). SpikEmo: Enhancing emotion recognition with spiking temporal dynamics in conversations. IEEE Transactions on Affective Computing, 13(3), 1201–1213. https://doi.org/10.1109/TAFFC.2021.3089257
[5] Luo, Y., Fu, Q., Xie, J., Qin, Y., Wu, G., Liu, J., Jiang, F., Cao, Y., & Ding, X. (2020). EEG-based emotion classification using spiking neural networks. Frontiers in Neuroscience, 14, 1030. https://doi.org/10.3389/fnins.2020.01030
[6] Alzhrani, W., Doborjeh, M., Doborjeh, Z., & Kasabov, N. (2021). Emotion recognition using EEG data in a brain-inspired spiking neural network architecture. In Proceedings of the International Conference on Neural Information Processing (ICONIP) (pp. 163–175).
[7] Mansouri-Benssassi, E., & Ye, J. (2021). Investigating multisensory integration in emotion recognition through bio-inspired computational models. Cognitive Computation, 13, 1253–1271. https://doi.org/10.1007/s12559-020-09796-x
[8] Wang, B., Dong, G., Zhao, Y., Li, R., Yang, H., Yin, W., & Liang, L. (2021). Spiking emotions: Dynamic vision emotion recognition using spiking neural networks. Pattern Recognition Letters, 145, 23–31. https://doi.org/10.1016/j.patrec.2021.02.001
[9] Mansouri-Benssassi, E., & Ye, J. (2021). Generalisation and robustness investigation for facial and speech emotion recognition using bio-inspired spiking neural networks. Cognitive Neurodynamics, 15(1), 85–99. https://doi.org/10.1007/s11571-020-09650-0
[10] Tan, C., Ceballos, G., Kasabov, N., & Subramaniyam, N. P. (2021). FusionSense: Emotion classification using feature fusion of multimodal data in brain-inspired spiking neural networks. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME) (pp. 432–437).
[11] Chunduri, R. K., & Perera, D. G. (2023). Neuromorphic sentiment analysis using spiking neural networks. ACM Transactions on Neuromorphic Systems, 2(1), 1–15.
[12] Jain, M., & Shukla, S. (2022). Accurate speech emotion recognition using brain-inspired decision-making spiking neural networks. Neural Processing Letters, 55(1), 1871–1895. https://doi.org/10.1007/s11063-021-10567-7
[13] Li, W., Fang, C., Zhu, Z., Chen, C., & Song, A. (2023). Fractal spiking neural network scheme for EEG-based emotion recognition. IEEE Transactions on Cognitive and Developmental Systems, 14(4), 807–819. https://doi.org/10.1109/TCDS.2022.3140231
[14] Barchid, S., Allaert, B., Aissaoui, A., Mennesson, J., & Djéraba, C. (2023). Spiking-FER: Spiking neural network for facial expression recognition with event cameras. Neurocomputing, 446, 112–128. https://doi.org/10.1016/j.neucom.2020.09.009
[15] Fu, S.-Y., Yang, G.-S., & Kuai, X.-K. (2021). A spiking neural network-based cortex-like mechanism and application to facial expression recognition. Frontiers in Neuroscience, 16, 645. https://doi.org/10.3389/fnins.2021.00645
[16] Cui, S., Lee, D., & Wen, D. (2024). Toward brain-inspired foundation model for EEG signal processing: our opinion. *Frontiers in Neuroscience*, *18*, 1507654.
[17] Alenizi, F., Bouallegue, B., Sam, A., & Boostani, R. (2024). EEG-Based Depression Classification and Brain Region Analysis Using a Hybrid of NeuCube and Dictionary Learning Framework. *Authorea Preprints*.
[18] https://analyticsindiamag.com/ai-trends/a-tutorial-on-spiking-neural-networks-for-beginners/
[19] Chen, S.Y., Hsu, C.C., Kuo, C.C. and Ku, L.W. EmotionLines: An Emotion Corpus of Multi-Party Conversations. arXiv preprint arXiv:1802.08379 (2018).
[20] Livingstone SR, Russo FA (2018) 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 ONE13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391
[21] Bird, Jordan & Ekart, Aniko & Buckingham, Christopher & Faria, Diego. (2019). Mental Emotional Sentiment Classification with an EEG-based Brain-machine Interface.