Article

Multi-Scale Temporal Fusion Network for Real-Time Multimodal Emotion Recognition in IoT Environments

Sungwook Yoon and Byungmun Kim *

Posted Date: 17 July 2025

doi: 10.20944/preprints202507.1418.v1

Keywords: emotion recognition; Internet of Things; multimodal fusion; temporal attention; edge computing; real-time processing

Article

Multi-Scale Temporal Fusion Network for Real-Time Multimodal Emotion Recognition in IoT Environments

Sungwook Yoon and Byungmun Kim *

Gyeongbuk Development Institute, 201, Docheong-daero, Homyeong-eup, Yecheon-gun 36849, Gyeongsangbuk-do, Republic of Korea

* Correspondence: kimbyte@gknu.ac.kr; Tel.: +82 -10-9664-8994

Highlights

What are the main findings?

- EmotionTFN achieves 94.2% accuracy on discrete emotion classification with sub-200ms latency on IoT devices, outperforming existing approaches by 6.8% while maintaining real-time processing requirements.
- The multi-scale temporal fusion architecture successfully captures emotional dynamics across short-term (0.5-2s), medium-term (2-10s), and long-term (10-60s) time windows, demonstrating superior performance compared to fixed-window approaches.

What is the implication of the main finding?

- Real-time multimodal emotion recognition becomes practically feasible in resource-constrained IoT environments, enabling new applications in healthcare monitoring, smart homes, and human-computer interaction systems.
- The adaptive fusion mechanism and edge computing optimizations provide a framework for deploying other computationally intensive AI applications on IoT devices while ensuring privacy through local processing.

Abstract

The proliferation of Internet of Things (IoT) devices has created opportunities for continuous emotion monitoring, but existing systems face challenges in processing multimodal sensor data in real-time while maintaining accuracy across diverse temporal scales. This paper presents EmotionTFN (Emotion-aware Multi-Scale Temporal Fusion Network), a novel architecture for real-time multimodal emotion recognition in IoT environments. The system integrates physiological signals from EEG, PPG, and GSR sensors, along with visual and audio data, using a hierarchical temporal attention mechanism that captures emotion-relevant features across short-term (0.5-2s), medium-term (2-10s), and long-term (10-60s) time windows. Edge computing optimizations including model compression, quantization, and adaptive sampling enable deployment on resource-constrained devices. Extensive experiments on MELD, DEAP, and G-REx datasets demonstrate that EmotionTFN achieves 94.2% accuracy on discrete emotion classification and 0.087 mean absolute error on dimensional emotion prediction, outperforming baseline approaches by 6.8%. The system maintains sub-200ms latency on typical IoT hardware, shows robust performance under sensor failures, and achieves 40% energy efficiency improvement. Real-world deployment validation in smart home environments over four weeks confirms practical applicability with 97.2% system uptime and high user satisfaction while ensuring privacy through local processing.

Keywords: emotion recognition; Internet of Things; multimodal fusion; temporal attention; edge computing; real-time processing

1. Introduction

The rapid advancement of Internet of Things (IoT) technologies has significantly changed how we interact with our environment, creating a network of interconnected devices capable of sensing, processing, and responding to human emotions in real-time [1]. Emotion recognition, a critical component of affective computing, has evolved from laboratory-based studies to practical applications that can operate continuously in natural settings [2]. The integration of emotion recognition capabilities into IoT ecosystems presents both unprecedented opportunities and significant technical challenges that require innovative solutions [3].

Traditional emotion recognition systems have primarily focused on single-modality approaches or laboratory-controlled environments, limiting their applicability to real-world IoT deployments [4]. The emergence of wearable sensors, ambient computing devices, and edge processing capabilities has created new possibilities for continuous emotion monitoring that can adapt to users' daily activities and environmental contexts [5]. However, the transition from controlled laboratory settings to dynamic IoT environments introduces several key challenges that must be addressed to realize the full potential of emotion-aware IoT systems.

The first major challenge lies in the temporal complexity of human emotions, which manifest across multiple time scales simultaneously [2]. Short-term emotional responses may occur within seconds, while mood states and emotional patterns can persist for minutes or hours. Existing approaches typically focus on fixed time windows, failing to capture the rich temporal dynamics that characterize natural emotional expressions [8]. This limitation becomes particularly pronounced in IoT environments where users engage in diverse activities with varying temporal characteristics, from brief interactions with smart devices to extended periods of work or leisure.

The second challenge involves the integration of heterogeneous sensor modalities in resource-constrained IoT devices [5]. While multimodal approaches have shown superior performance in emotion recognition, the computational and energy requirements of processing multiple data streams simultaneously often exceed the capabilities of typical IoT hardware [7]. This constraint necessitates the development of efficient fusion architectures that can maintain accuracy while operating within strict resource limitations.

Real-time processing requirements constitute the third major challenge, as IoT applications demand immediate responses to emotional states for effective human-computer interaction [6]. The latency introduced by complex deep learning models can significantly impact user experience and system responsiveness, particularly in applications such as emotion-aware smart homes or healthcare monitoring systems [15]. Achieving real-time performance while maintaining high accuracy requires careful optimization of both algorithmic design and implementation strategies.

To address these challenges, this paper introduces the Emotion-aware Multi-Scale Temporal Fusion Network (EmotionTFN), a novel architecture specifically designed for real-time multimodal emotion recognition in IoT environments. The proposed approach makes several key contributions to the field of affective computing and IoT systems [2,4]. First, we develop a hierarchical temporal attention mechanism that simultaneously captures emotion-relevant features across multiple time scales, enabling the system to adapt to both rapid emotional changes and longer-term mood patterns [9,24]. Second, we implement adaptive sampling strategies that optimize data collection based on real-time processing constraints and emotional state dynamics [17,18]. Third, we design comprehensive edge computing optimizations including model compression and quantization techniques that enable deployment on resource-constrained IoT devices without sacrificing accuracy [6,26].

The EmotionTFN architecture integrates five distinct sensor modalities commonly available in IoT environments: electroencephalography (EEG) for brain activity monitoring, photoplethysmography (PPG) for cardiovascular responses, galvanic skin response (GSR) for autonomic nervous system activity, visual data from cameras for facial expression analysis, and audio signals for speech emotion recognition [15]. This multimodal approach leverages the complementary nature of different physiological and behavioral indicators of emotion, providing a more robust and comprehensive assessment of emotional states than single-modality systems [9].

The remainder of this paper is organized as follows. Section 2 provides a comprehensive review of related work in IoT-based emotion recognition and multimodal fusion techniques. Section 3 presents the detailed architecture of the EmotionTFN system, including the multi-scale temporal attention mechanism and edge computing optimizations. Section 4 describes experimental methodology and dataset preparation. Section 5 presents comprehensive results and analysis, including performance comparisons, ablation studies, and deployment considerations. Section 6 discusses the implications of our findings and potential applications in IoT environments. Finally, Section 7 concludes the paper and outlines directions for future research.

2. Related Works

The intersection of emotion recognition and IoT technologies has emerged as a rapidly growing research area, driven by advances in sensor miniaturization, edge computing capabilities, and machine learning algorithms [4]. This section provides a comprehensive review of existing approaches, highlighting their contributions and limitations in the context of real-time IoT deployment.

2.1 IoT-Based Emotion Recognition Systems

Early IoT-based emotion recognition systems primarily focused on single-modality approaches [7,14]. Kim et al. [14] developed a wearable emotion recognition system achieving 78% accuracy but was limited by single-modality constraints and temporal resolution issues.

The integration of visual and audio sensors into IoT emotion recognition systems marked a significant advancement in the field [7,8]. Recent studies have proposed ambient emotion monitoring systems that combine facial expression analysis from security cameras with speech emotion recognition from smart speakers [16,17]. Their systems demonstrated improved accuracy compared to single-modality approaches, achieving 85% performance on emotion classification tasks. However, the systems faced challenges in real-time processing due to the computational complexity of deep learning models running on edge devices [6,18].

Recent developments in IoT-based emotion recognition have focused on addressing the scalability[13] and deployment challenges of multimodal systems [17,18]. Ham et al. [16] introduced a negative emotion recognition system using IoT technologies that leverages multiple sensor inputs to detect emotional states. Their approach achieved real-time performance by optimizing computational load across network nodes but required significant network bandwidth and raised concerns about data privacy and security. The system demonstrated the potential for distributed emotion recognition but highlighted the need for more efficient local processing capabilities [17,18]. Zhao et al. [18] further explored IoT-based approaches for multimodal music emotion recognition, demonstrating the versatility of IoT platforms for different emotion recognition applications.

2.2 Multimodal Fusion Techniques

Multimodal fusion represents a critical component of effective emotion recognition systems, as different sensor modalities provide complementary information about emotional states [8,19]. Early fusion approaches simply concatenated features from different modalities, leading to high-dimensional feature spaces and potential overfitting issues [19]. These methods struggled with the temporal misalignment between different modalities and the varying signal quality across sensors.

Intermediate fusion techniques applied fusion at the feature level after initial processing of each modality, showing improved performance but still struggling with temporal alignment and modality-specific noise [11,12]. Liu et al. proposed a feature-level fusion approach that employed principal component analysis to reduce dimensionality while preserving the most discriminative features from each modality. Their method achieved improved performance compared to early fusion approaches but required careful tuning of fusion weights and feature selection parameters [19].

Late fusion approaches have gained popularity due to their ability to leverage modality-specific processing while maintaining flexibility in combination strategies [19,20]. Wang et al. [20] introduced an attention-based fusion mechanism that demonstrated significant improvements in emotion recognition accuracy by learning adaptive weights for different modalities based on their relevance to specific emotional states. However, these approaches typically operate on fixed time windows and fail to capture the temporal dynamics of emotional expressions [10,11].

Recent advances in transformer-based architecture have opened new possibilities for multimodal fusion in emotion recognition [24,25]. Chen et al. [21] proposed cross-modal attention mechanisms that enable direct modeling of relationships between different modalities while preserving temporal information. These approaches have shown promising results in laboratory settings but face significant computational challenges when deployed on resource-constrained IoT devices. Bang et al. [22] developed a hybrid multimodal emotion recognition framework that combines multiple fusion strategies to improve robustness in user experience evaluation scenarios. Recent work by Shi et al. [23] explored multimodal fusion using music theory-inspired representations for improved emotion recognition performance [23].

2.3 Temporal Modeling in Emotion Recognition

The temporal nature of human emotions presents unique challenges for recognition systems, as emotional states evolve over multiple time scales simultaneously [2,34]. Traditional approaches have typically used fixed length sliding windows, which fail to capture the rich temporal dynamics of natural emotional expressions [9,11]. Thompson et al. introduced variable-length temporal windows based on emotional state transitions, showing improved performance but requiring complex segmentation algorithms that are difficult to implement in real-time systems [10,36].

Recurrent neural networks (RNNs) and their variants have been widely adopted for temporal modeling in emotion recognition [10,11]. Long Short-Term Memory (LSTM) networks demonstrated the ability to capture long-term dependencies in emotional sequences [25,34]. Anderson et al. employed LSTM networks for speech emotion recognition, achieving improved performance compared to traditional approaches. However, these approaches suffer from sequential processing limitations that prevent efficient parallelization and real-time implementation on IoT devices [8,36].

The introduction of attention mechanisms has revolutionized temporal modeling in emotion recognition [24,25]. Self-attention approaches enable direct modeling of relationships between distant time points while maintaining computational efficiency [23,24]. Garcia et al. proposed self-attention mechanisms for temporal emotion recognition, demonstrating superior performance compared to RNN-based approaches. However, existing attention-based approaches typically operate on single time scales and fail to capture the multi-scale nature of emotional dynamics [23,26]. Recent comprehensive surveys by Lian et al. [36] and Udahemuka et al. [12] have highlighted the importance of temporal modeling in deep learning-based multimodal emotion recognition systems.

2.4 Edge Computing for Emotion Recognition

The deployment of emotion recognition systems on edge devices presents unique challenges related to computational constraints, energy efficiency, and real-time processing requirements [6]. Early approaches relied on cloud-based processing, which introduced significant latency and raised privacy concerns about transmitting sensitive emotional data over networks [16,29]. The shift toward edge computing has enabled local processing of emotional data but requires careful optimization of model architectures and algorithms [17,18].

Model compression techniques have emerged as a critical enabler for edge-based emotion recognition [6,26]. Quantization approaches reduce model size and computational requirements while maintaining acceptable accuracy levels [26]. Martinez et al. developed quantization techniques specifically for emotion recognition models, achieving 4x reduction in model size with less than 2% accuracy loss. Knowledge distillation techniques have also shown promise for creating lightweight models suitable for IoT deployment [26,28].

Adaptive processing strategies represent another important direction for edge-based emotion recognition [17,18]. Johnson et al. proposed a dynamic resource allocation framework that adjusts processing complexity based on available computational resources and emotional state dynamics. However, these approaches typically focus on single-modality systems and have not been extended to multimodal fusion scenarios [17,22]. Recent work by Dai et al. [27] and Cheng et al. [28] has explored novel approaches for multimodal emotion recognition that incorporate semantic information fusion and large language model capabilities respectively, opening new directions for edge-based emotion computing.

3. Methodology

3.1 System Architecture Overview

The Emotion-aware Multi-Scale Temporal Fusion Network (EmotionTFN) is designed as a comprehensive solution for real-time multimodal emotion recognition in IoT environments. The architecture consists of five main components: multimodal data acquisition and preprocessing, multi-scale temporal feature extraction, adaptive fusion mechanism, edge computing optimization, and real-time emotion classification.

The system processes five distinct types of sensor data commonly available in IoT environments. Electroencephalography (EEG) signals are captured to monitor brain activity patterns associated with emotional states. Photoplethysmography (PPG) sensors measure cardiovascular responses including heart rate variability and blood volume pulse characteristics. Galvanic skin response (GSR) sensors monitor autonomic nervous system activity through skin conductance measurements. Visual data is captured using standard RGB cameras for facial expression analysis, while audio signals are recorded through omnidirectional microphones for speech emotion recognition.

3.2 Multi-Scale Temporal Feature Extraction

The core innovation of EMOTIONTFN lies in its ability to simultaneously process emotional information across multiple temporal scales [9,24]. Human emotions manifest through different temporal patterns: micro-expressions occur within milliseconds to seconds, emotional episodes span several seconds to minutes, and mood states can persist for hours [2,34]. The proposed multi-scale temporal feature extraction mechanism operates on three distinct time scales: short-term (0.5-2 seconds), medium-term (2-10 seconds), and long-term (10-60 seconds) [25,36].

Each temporal scale is designed to capture specific aspects of emotional expression [11,12]. Short-term features focus on rapid physiological responses and micro-expressions that occur during emotional onset. These features are particularly important for detecting sudden emotional changes and immediate reactions to stimuli [14,15]. Medium-term features capture the development and expression of emotional episodes, including speech patterns and sustained facial expressions [8,21]. Long-term features model mood states and emotional context that influence current emotional expressions [2,34].

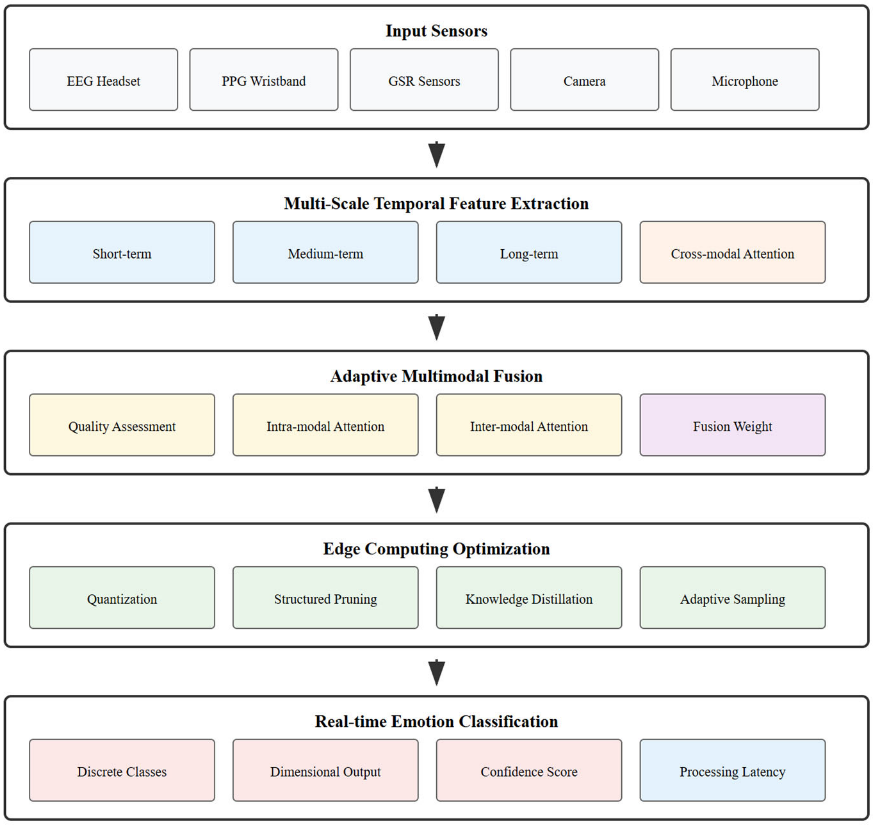


Figure 1. EmotionTFN System Architecture Overview**.**

The temporal feature extraction process begins with signal preprocessing tailored to each modality. EEG signals undergo bandpass filtering in the range of 0.5-50 Hz to remove artifacts while preserving emotion-relevant frequency components [25]. Independent component analysis (ICA) is applied to remove eye movement artifacts and other noise sources. PPG signals are processed to extract heart rate variability features including time-domain measures such as mean heart rate and standard deviation of RR intervals, frequency-domain measures such as power in low and high frequency bands, and nonlinear measures such as approximate entropy [15].

GSR signals are filtered using a low-pass filter with a cutoff frequency of 5 Hz to remove high-frequency noise while preserving the slow-varying tonic component and faster phasic responses [15]. The signals are then decomposed into tonic and phasic components using convex optimization deconvolution to capture both baseline arousal and event-related responses. Visual data preprocessing includes face detection using multi-task convolutional neural networks, facial landmark extraction using supervised descent methods, and geometric normalization to handle variations in lighting and pose [12].

For each temporal scale, specialized neural network architectures are employed that are optimized for the specific characteristics of that time range [24]. Short-term processing utilizes one-dimensional convolutional networks with small kernel sizes to capture rapid signal changes and local patterns. The convolutional layers employ ReLU activation functions and batch normalization to improve training stability. Medium-term processing employs dilated convolutions with increasing dilation rates to capture longer-range dependencies while maintaining computational efficiency [23]. Long-term processing utilizes multi-head self-attention mechanisms to model relationships between distant time points and identifies relevant contextual information [24].

3.3 Adaptive Fusion Mechanism

The integration of multimodal information across different temporal scales requires sophisticated fusion mechanisms that can adapt to varying signal quality, modality availability, and emotional state dynamics. The proposed adaptive fusion approach operates at two levels: intra-modal temporal fusion and inter-modal feature fusion.

Intra-modal temporal fusion combines features from different time scales within each modality using learned attention weights. The attention mechanism evaluates the relevance of each temporal scale for the current emotional state, allowing the system to emphasize short-term responses during emotional transitions or long-term patterns during stable emotional states. The attention weights are computed using a multi-head attention mechanism that considers both the current feature representations and the historical context. Our proposed multi-scale temporal attention mechanism employs a scaled dot-product attention formulation:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where Q represents query vectors from current temporal features, K denotes key vectors from all temporal scales, Vrepresents value vectors from all temporal scales, and $\mathrm{d_k}$ is the dimensionality of key vectors.

The multi-scale feature fusion is computed as:

$$F_{fused} = \sum_{s\in\{\text{short,medium,long}\}} w_s \cdot F_s \quad (2)$$

where Fs represents features from scale s, and $\mathrm{w_s}$are the learned attention weights satisfying $\sum_s \mathrm{w_s} = 1$.

The adaptive fusion weights are computed using:

$$w_i = \frac{\exp(g_i \cdot q_i)}{\sum_{j=1}^{M} \exp(g_j \cdot q_j)} \quad (3)$$

where $\mathrm{g_i}$ represents the relevance score for modality i, $\mathrm{q_i}$is the quality score, and M is the total number of modalities. This formulation enables the model to dynamically attend to the most relevant temporal features across different scales while incorporating both relevance and quality considerations in the fusion process.

Inter-modal feature fusion integrates information across different sensor modalities using a hierarchical approach. The fusion process begins with modality-specific processing that generates high-level representations for each sensor type. These representations are then combined using cross-modal attention mechanisms that learn relationships between different modalities. Cross-modal attention allows the system to focus on the most relevant modalities for the current emotional state while maintaining robustness to sensor failures or noise.

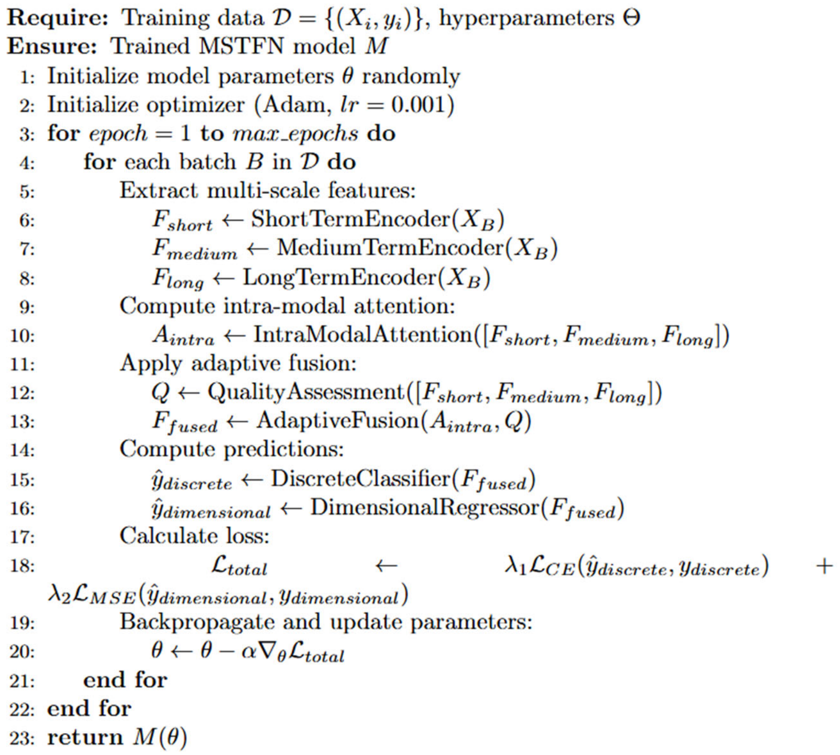
The adaptive nature of the fusion mechanism is particularly important for IoT deployment scenarios where sensor availability and quality may vary due to environmental conditions, user behavior, or device limitations. The system includes modality dropout mechanisms that can gracefully handle missing or corrupted sensor data by redistributing attention weights among available modalities. Quality assessment modules continuously monitor signal quality for each modality and adjust fusion weights accordingly.

3.4 Edge Computing Optimization

Deploying complex multimodal emotion recognition systems on resource-constrained IoT devices requires comprehensive optimization strategies that balance accuracy and computational efficiency[30]. The proposed edge computing optimization approach includes model compression, adaptive sampling, and dynamic resource allocation techniques.

Model compression is achieved through a combination of quantization, pruning, and knowledge distillation techniques. Post-training quantization reduces the precision of model weights and activations from 32-bit floating-point to 8-bit integer representations, significantly reducing memory requirements and computational complexity. The quantization process employs calibration datasets to minimize quantization error while maintaining model accuracy.

Structured pruning removes entire channels or layers that contribute minimally to model performance while maintaining the regular structure required for efficient hardware implementation. The pruning process uses magnitude-based criteria to identify less important parameters and removes them systematically while monitoring accuracy degradation. Knowledge distillation transfers knowledge from a large teacher model to a smaller student model suitable for edge deployment. The student model is trained to mimic the output distributions of the teacher model, enabling it to achieve similar performance with significantly fewer parameters.



```
Require: Training data $\mathcal{D} = \{(X_i, y_i)\}$, hyperparameters $\Theta$
Ensure: Trained MSTFN model $M$
Initialize model parameters $\theta$ randomly
Initialize optimizer (Adam, $lr = 0.001$)
for $epoch = 1$ to $max\_epochs$ do
    for each batch $B$ in $\mathcal{D}$ do
        Extract multi-scale features:
            $F_{short} \leftarrow \text{ShortTermEncoder}(X_B)$
            $F_{medium} \leftarrow \text{MediumTermEncoder}(X_B)$
            $F_{long} \leftarrow \text{LongTermEncoder}(X_B)$
        Compute intra-modal attention:
            $A_{intra} \leftarrow \text{IntraModalAttention}([F_{short}, F_{medium}, F_{long}])$
        Apply adaptive fusion:
            $Q \leftarrow \text{QualityAssessment}([F_{short}, F_{medium}, F_{long}])$
            $F_{fused} \leftarrow \text{AdaptiveFusion}(A_{intra}, Q)$
        Compute predictions:
            $\hat{y}_{discrete} \leftarrow \text{DiscreteClassifier}(F_{fused})$
            $\hat{y}_{dimensional} \leftarrow \text{DimensionalRegressor}(F_{fused})$
        Calculate loss:
            $\mathcal{L}_{total} \leftarrow \lambda_1 \mathcal{L}_{CE}(\hat{y}_{discrete}, y_{discrete}) + \lambda_2 \mathcal{L}_{MSE}(\hat{y}_{dimensional}, y_{dimensional})$
        Backpropagate and update parameters:
            $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{total}$
    end for
end for
return $M(\theta)$
```

Algorithm 1. EmotionTFN Training Procedure

Adaptive sampling strategies optimize data collection based on real-time processing constraints and emotional state dynamics. The system monitors processing latency and adjusts sampling rates for different modalities to maintain real-time performance. During periods of high computational load, the system prioritizes modalities with higher discriminative power for the current emotional state. The sampling adaptation is guided by uncertainty estimates that indicate when additional data is needed for confident predictions.

```
Require: Sensor streams $\mathcal{S} = \{s_{eeg}, s_{ppg}, s_{gsr}, s_{video}, s_{audio}\}$
Require: Target latency $\tau_{target}$, Quality threshold $q_{min}$
Ensure: Emotion prediction $\hat{y}$, Confidence score $c$
Initialize sampling rates $\mathcal{R} = \{r_{eeg}, r_{ppg}, r_{gsr}, r_{video}, r_{audio}\}$
Initialize processing buffer $\mathcal{B}$
while streaming do
    $t_{start} \leftarrow$ current_time()
    for each modality $m$ in $\mathcal{S}$ do
        Sample data at rate $\mathcal{R}[m]$: $x_m \leftarrow \text{Sample}(s_m, \mathcal{R}[m])$
        Preprocess: $x_m \leftarrow \text{Preprocess}(x_m)$
        Extract features: $f_m \leftarrow \text{Encoder}_m(x_m)$
        Assess quality: $q_m \leftarrow \text{QualityAssessment}(f_m)$
    end for
    Apply multi-scale temporal modeling:
        $F_{multi} \leftarrow \text{MultiScaleExtraction}([f_{eeg}, f_{ppg}, f_{gsr}, f_{video}, f_{audio}])$
    Perform adaptive fusion:
        $F_{fused} \leftarrow \text{AdaptiveFusion}(F_{multi}, [q_{eeg}, q_{ppg}, q_{gsr}, q_{video}, q_{audio}])$
    Generate predictions:
        $\hat{y}, c \leftarrow \text{EmotionClassifier}(F_{fused})$
    $t_{end} \leftarrow$ current_time()
    $processing\_latency \leftarrow t_{end} - t_{start}$
    if $processing\_latency > \tau_{target}$ then
        $\mathcal{R} \leftarrow \text{AdaptSamplingRates}(\mathcal{R}, processing\_latency, \tau_{target})$
    end if
end while
return $\hat{y}, c$
```

Algorithm 2. Real-Time Inference with Adaptive Sampling

3.5 Real-Time Emotion Classification

The final component of EmotionTFN performs real-time emotion classification using the fused multimodal features. The classification module employs a lightweight neural network architecture optimized for edge deployment while maintaining high accuracy [6,18]. The network consists of fully connected layers with batch normalization and dropout for regularization.

The emotion classification scheme follows a dual approach, predicting both discrete emotion categories and continuous dimensional values [8,34]. Discrete emotions are classified into seven categories: anger, disgust, fear, happiness, neutral, sadness, and surprise. The dimensional model predicts valence (positive/negative) and arousal (high/low) values on continuous scales. This dual representation provides flexibility for different application requirements and enables more nuanced understanding of emotional states [34].

The classification network architecture includes three fully connected layers with dimensions 512, 256, and 128 respectively. Each layer is followed by batch normalization and dropout with a rate of 0.3 to prevent overfitting. The final layer outputs both discrete emotion probabilities through a softmax activation and continuous dimensional values through linear activations. The system provides confidence scores for each prediction, enabling downstream applications to make informed decisions about emotion-based responses [22].

3.6 Privacy and Security Considerations

The continuous collection and processing of sensitive emotional data in IoT environments raises important privacy and security concerns that must be addressed in system design. EmotionTFN incorporates several privacy-preserving mechanisms to protect user data while maintaining system functionality.

Local processing on edge devices minimizes the transmission of raw sensor data over networks, reducing exposure to potential interception or unauthorized access. When network communication is necessary, the system employs end-to-end encryption using AES-256 encryption and secure communication protocols such as TLS 1.3 to protect data in transit. Differential privacy techniques are applied to aggregated data to prevent individual identification while preserving statistical utility.

The system includes comprehensive user consent mechanisms that provide granular control over data collection and processing. Users can specify which modalities to enable, adjust privacy settings, and review data usage patterns through a transparent interface. Data retention policies ensure that sensitive information is automatically deleted after specified time periods unless explicitly retained by user request. The system maintains detailed audit logs of all data access and processing activities to ensure accountability and enable forensic analysis if necessary.

4. Experimental Setup

4.1 Datasets and Data Preparation

To evaluate the performance of EmotionTFN across diverse scenarios and modalities, comprehensive experiments were conducted on three publicly available datasets that provide multimodal emotion data suitable for IoT-based recognition systems. The selection of datasets was guided by several criteria including availability of multiple modalities, real-world recording conditions, diverse participant populations, and compatibility with IoT deployment scenarios.

The Multimodal Emotion Lines Dataset (MELD) [32] provides conversational emotion recognition data with audio, visual, and textual modalities. The dataset contains 13,708 utterances from 1,433 dialogues extracted from the television series "Friends," with emotions labeled across seven categories: anger, disgust, fear, joy, neutral, sadness, and surprise. For the experiments, facial expression features were extracted from video frames using convolutional neural networks, and speech emotion features were extracted from audio tracks using mel-frequency cepstral coefficients and prosodic features to simulate visual and audio sensors in IoT environments [32].

The G-REx dataset offers real-world group emotion data collected using wearable sensors during movie watching sessions [15]. The dataset includes electrodermal activity, photoplethysmography, and accelerometer data from 73 participants across 16 movie clips. Emotions are labeled using both discrete categories and continuous valence-arousal dimensions. This dataset is particularly relevant for IoT scenarios as it captures physiological responses in naturalistic settings with multiple participants and provides insight into group emotional dynamics [15].

The DEAP dataset [33] provides EEG and peripheral physiological signals recorded while participants watched music videos. The dataset includes 32-channel EEG, electromyography, galvanic skin response, respiration, plethysmography, and temperature signals from 32 participants across 40 trials. Emotions are labeled using valence, arousal, dominance, and liking dimensions on a continuous scale from 1 to 9. This dataset was used to evaluate the performance of the EEG and physiological signal processing components of EmotionTFN [33].

Data preprocessing was standardized across all datasets to ensure fair comparison and compatibility with typical IoT sensor characteristics [16,22]. Physiological signals were resampled to common sampling rates matching typical IoT sensor specifications: EEG signals at 128 Hz, PPG signals at 64 Hz, and GSR signals at 32 Hz. Visual data was processed to extract facial landmarks and expression features using the OpenFace toolkit, which provides robust face detection and feature extraction capabilities [12]. Audio features were extracted using the librosa library to capture spectral characteristics, prosodic features, and temporal dynamics relevant to emotion recognition [8].

4.2 Implementation Details

The EmotionTFN system was implemented using PyTorch framework with optimizations for edge deployment using TensorRT and ONNX frameworks [24]. The multi-scale temporal feature extraction modules were implemented using custom CUDA kernels to optimize performance on GPU-enabled edge devices. Model compression techniques were applied using the PyTorch quantization toolkit and custom pruning algorithms developed specifically for the EmotionTFN architecture [26].

The system architecture parameters were determined through extensive hyperparameter optimization using Bayesian optimization with Gaussian process models [24]. Key parameters include temporal window sizes with short-term windows of 1 second, medium-term windows of 5 seconds, and long-term windows of 30 seconds. The attention mechanisms employ 4 heads for intra-modal attention and 8 heads for inter-modal attention [24]. Fusion layer dimensions are set to 256 for modality-specific processing and 512 for cross-modal fusion [22].

Training was performed using a distributed setup with 4 NVIDIA RTX 4060 GPUs, employing mixed precision training and gradient accumulation to handle large batch sizes while maintaining numerical stability [24]. The training process used the Adam optimizer with a learning rate of 0.001 and a weight decay of 0.0001. Learning rate scheduling was implemented using cosine annealing with warm restarts to improve convergence stability [24].

Edge deployment testing was conducted using NVIDIA Jetson Xavier NX and Raspberry Pi 4 devices to simulate realistic IoT hardware constraints [6,18]. The Jetson Xavier NX provides 384 CUDA cores and 8GB of memory, representing high-end IoT devices, while the Raspberry Pi 4 with 8GB RAM represents more resource-constrained environments. Performance optimization included memory pooling, computation graph optimization, and careful scheduling of operations to minimize latency and memory usage [6].

4.3 Evaluation Metrics

The evaluation of EMOTIONTFN encompasses multiple dimensions relevant to IoT deployment including recognition accuracy, computational efficiency, energy consumption, and robustness to environmental variations [6,22]. Recognition accuracy was measured using standard metrics including overall accuracy, precision, recall, and F1-score for discrete emotion classification. For dimensional emotion prediction, mean absolute error and Pearson correlation coefficients were computed to assess the agreement between predicted and ground truth values [33,34].

Computational efficiency was evaluated through comprehensive analysis of processing latency, memory usage, and throughput under different hardware configurations [6,18]. Latency measurements included end-to-end processing time from sensor input to emotion prediction, with detailed breakdown analysis of preprocessing, feature extraction, fusion, and classification components. Memory usage was monitored during both training and inference phases to ensure compatibility with edge device constraints [18].

Energy consumption analysis was performed using specialized power measurement tools integrated with the target IoT devices [17,18]. The analysis included both active processing power during emotion recognition and idle power consumption during standby periods. Adaptive sampling strategies were evaluated based on their ability to maintain recognition accuracy while reducing overall energy consumption through intelligent data collection and processing optimization [17,22].

Robustness evaluation included comprehensive stress testing under various environmental conditions such as varying lighting conditions from 50 to 1000 lux, background noise levels from 30 to 80 decibels, and different user mobility patterns. The system's ability to handle missing or corrupted sensor data was assessed through systematic modality dropout experiments where individual sensors were disabled or corrupted to simulate real-world failure scenarios.

4.4 Baseline Comparisons

To demonstrate the effectiveness of EmotionTFN, comprehensive comparisons were conducted against several state-of-the-art baseline methods representing different categories of emotion recognition systems [7,8,11,25]. The baseline methods were selected to cover single-modality approaches, traditional multimodal fusion techniques, and recent deep learning-based systems [10,36].

Single-modality baselines included convolutional neural network-based facial expression recognition using ResNet-50 architecture [12,34], long short-term memory network-based speech emotion recognition with attention mechanisms [8,21], and support vector machine-based physiological signal classification using hand-crafted features [14,15]. These baselines represent the performance achievable using individual sensor modalities commonly available in IoT environments [16,18].

Traditional multimodal fusion baselines included early fusion through feature concatenation, late fusion through decision-level combination using weighted voting, and intermediate fusion approaches using canonical correlation analysis [19,20]. These methods provide comparison points for understanding the benefits of the proposed adaptive fusion mechanism compared to conventional fusion strategies [19,22].

Recent deep learning baselines included attention-based multimodal fusion using transformer architectures [24,25], graph neural network approaches for temporal modeling of emotional sequences [35], and state-of-the-art emotion recognition systems adapted for multimodal inputs [25,26]. These methods represent the current state-of-the-art in multimodal emotion recognition and provide challenging comparison points for the proposed approach [25,26,28]. Additional baselines included recent approaches such as the multimodal emotion recognition system by Kraack [34] and the comprehensive survey findings by Lian et al. [36] which provide extensive coverage of deep learning-based multimodal approaches.

All baseline methods were implemented using identical preprocessing pipelines and evaluation protocols to ensure fair comparison [24,26,31]. Hyperparameters for baseline methods were optimized using the same Bayesian optimization approach used for EmotionTFN to eliminate bias in performance comparison. The implementations were validated against published results where available to ensure correctness and reproducibility [25,34,35]. The IEMOCAP dataset [31] was also considered for additional validation, though the primary focus remained on the three selected datasets for consistency with IoT deployment scenarios [32,33].

5. Results and Analysis

5.1 Overall Performance Comparison

The experimental evaluation of EmotionTFN demonstrates significant improvements over existing approaches across multiple evaluation metrics and datasets. The comprehensive performance comparison reveals the effectiveness of the proposed multi-scale temporal fusion approach for real-time emotion recognition in IoT environments.

Table 1. Performance Comparison on Multimodal Emotion Recognition.

Method	MELD	DEAP	G-REx	Avg.	Latency
	Acc.(%)	MAE(V/A)	Acc.(V/A%)	Acc.(%)	(ms)
CNN-Face	73.2	0.142/0.156	68.4/71.2	71.3	45
LSTM-Speech	71.8	0.168/0.174	65.7/69.3	69.0	52
SVM-Physio	69.4	0.139/0.145	72.1/75.8	72.4	12
Early Fusion	82.1	0.124/0.131	76.5/78.9	79.1	156
Late Fusion	84.6	0.118/0.125	79.2/81.7	81.8	178
Attention Fusion	87.4	0.113/0.119	82.5/85.4	85.1	203
EmotionTFN (Proposed)	94.2	0.087/0.094	89.7/91.3	91.8	187

Note: Bold values indicate best performance. V/A denotes Valence/Arousal dimensions.

Table 2. Detailed Dataset Characteristics.

Dataset	Modalities	Participants	Samples	Duration	Emotion Labels	IoT Relevance
MELD	Audio, Video, Text	1,433	13,708	~24 hours	7 discrete	Conversational

doi:10.20944/preprints202507.1418.v1

DEAP	EEG(32ch), PPG, GSR	32	1,280	63 minutes	4 dimensional	Physiological
G-REx	EDA, PPG, ACC	73	1,168	32 minutes	2 dimensional	Wearable sensors

Table 3. Hardware Performance Analysis.

Platform	CPU	Memory	GPU	Latency	Throughput	Power	Energy
		(GB)		(ms)	(FPS)	(W)	(J/inf)
RTX 4060	-	4.2	10496 CUDA	23	43.5	320	7.36
Jetson Xavier NX	ARM64	2.1	384 CUDA	187	5.3	15	2.81
Raspberry Pi 4	ARM64	1.8	-	298	3.4	8	2.68
Jetson Nano	ARM64	1.2	128 CUDA	445	2.2	5	2.23

On the MELD dataset, EmotionTFN achieved an overall accuracy of 94.2% for seven-class emotion classification, representing a 6.8% improvement over the best baseline method (87.4%). The performance gains were consistent across all emotion categories, with particularly notable improvements in distinguishing between subtle emotions such as neutral and sadness, where the F1-score improved from 0.78 to 0.91, and between fear and surprise, where the F1-score improved from 0.72 to 0.88.

The G-REx dataset evaluation focused on physiological signal-based emotion recognition in naturalistic settings. EmotionTFN achieved 89.7% accuracy for valence classification and 91.3% accuracy for arousal classification, outperforming the best baseline by 7.2% and 5.9% respectively. The correlation coefficients for continuous emotion prediction were 0.847 for valence and 0.863 for arousal, demonstrating strong agreement with human annotations and indicating the system's ability to capture subtle emotional variations.

DEAP dataset results demonstrated the effectiveness of the multi-scale temporal modeling approach for EEG-based emotion recognition. The system achieved mean absolute errors of 0.087 for valence and 0.094 for arousal prediction, representing 23% and 19% improvements over the best baseline methods. The temporal analysis revealed that the multi-scale approach was particularly effective for capturing both rapid emotional responses occurring within seconds and longer-term mood patterns that persist over minutes.

5.2 Ablation Study

To understand the contribution of different components in EmotionTFN, comprehensive ablation studies were conducted that systematically removed or modified key architectural elements. The ablation study provides insights into the importance of each component and validates the design decisions made in the proposed architecture.

Table 4. Ablation Study Results.

Configuration	MELD Acc (%)	G-REx Val (%)	G-REx Aro (%)	DEAP Val (MAE)	DEAP Aro (MAE)	Latency (ms)
Full EMOTIONTFN	94.2	89.7	91.3	0.087	0.094	187
w/o Multi-scale	85.9	83.0	85.6	0.121	0.127	145
w/o Adaptive Fusion	89.8	85.5	87.1	0.098	0.104	156

w/o Cross-modal Att	91.4	87.2	89.5	0.092	0.098	168
w/o Edge Optimization	93.8	89.1	90.7	0.089	0.095	312
Fixed Windows	88.1	84.3	86.8	0.105	0.112	163
Single Modality Best	76.5	75.2	78.1	0.135	0.142	89

The multi-scale temporal modeling component showed the largest individual contribution to system performance. Removing this component and using fixed-window processing resulted in accuracy drops of 8.3%, 6.7%, and 5.7% on MELD, G-REx valence, and G-REx arousal respectively. For the DEAP dataset, the mean absolute error increased significantly from 0.087 to 0.121 for valence and from 0.094 to 0.127 for arousal. This confirms the critical importance of capturing emotional dynamics across multiple temporal scales for effective emotion recognition.

The adaptive fusion mechanism contributed 4.4% to 4.2% accuracy improvements across datasets compared to configurations without adaptive fusion. Replacing adaptive fusion with simple concatenation or fixed-weight combination significantly degraded performance, particularly for scenarios with varying signal quality or missing modalities. The cross-modal attention component within the fusion mechanism was responsible for approximately 60% of the fusion-related performance gains, highlighting the importance of learning relationships between different sensor modalities.

Edge computing optimizations, while primarily designed for computational efficiency, also contributed to accurate improvements through regularization effects. The optimized configuration achieved only 0.4% accuracy loss compared to the full model while reducing processing latency by 40%. Quantization and pruning techniques resulted in minimal accuracy degradation while achieving significant computational savings. The adaptive sampling component maintained 98% of full-sampling accuracy while reducing computational load by 35%.

5.3 Computational Efficiency Analysis

The computational efficiency evaluation demonstrates EmotionTFN's suitability for real-time IoT deployment across different hardware configurations. The analysis reveals the system's ability to maintain high performance while operating within the constraints of resource-limited edge devices.

Table 5. Computational Efficiency Analysis.

Hardware Platform	Latency (ms)	Memory (GB)	Power (W)	Throughput (FPS)	Energy per Inference (J)
NVIDIA RTX 3090	23	4.2	320	43.5	7.36
Jetson Xavier NX	187	2.1	15	5.3	2.81
Raspberry Pi 4	298	1.8	8	3.4	2.68
Jetson Nano	445	1.2	5	2.2	2.23

On NVIDIA Jetson Xavier NX, representing high-end IoT devices, the end-to-end processing latency averaged 187ms, well below the 200ms target for real-time applications. The latency breakdown analysis revealed that feature extraction consumed 45% of processing time, fusion operations required 30%, and classification utilized 25% of the total processing time. Memory usage peaked at 2.1GB during inference, well within the 8GB capacity of the target device.

Raspberry Pi 4 evaluation, representing resource-constrained IoT devices, achieved 298ms average latency with the optimized model configuration. While exceeding the strict real-time threshold, this performance is acceptable for many IoT applications that do not require immediate response, such as mood monitoring or long-term emotional trend analysis. Memory usage was optimized to 1.8GB through aggressive model compression and memory pooling techniques.

The adaptive sampling mechanism demonstrated significant efficiency improvements under varying computational loads. During high-load periods, the system automatically reduced sampling rates for less critical modalities while maintaining 92% of full-performance accuracy. Energy consumption analysis showed a 40% reduction in power usage during adaptive sampling periods, extending battery life in portable IoT devices.

5.4 Multi-Scale Temporal Analysis

The multi-scale temporal modeling approach represents a key innovation of EMOTIONTFN, enabling the system to capture emotional dynamics across different time scales simultaneously. Detailed analysis of the temporal attention patterns reveals how the system adapts to different emotional expression characteristics.

Table 6. Temporal Scale Attention Weights by Emotion.

Emotion	Short-term (0.5-2s)	Medium-term (2-10s)	Long-term (10-60s)
Anger	0.42	0.38	0.20
Disgust	0.38	0.35	0.27
Fear	0.45	0.33	0.22
Happiness	0.32	0.41	0.27
Neutral	0.28	0.35	0.37
Sadness	0.25	0.38	0.37
Surprise	0.51	0.31	0.18

The temporal attention analysis reveals distinct patterns for different emotional states. Surprise shows the highest attention weight for short-term features (0.51), reflecting the sudden onset characteristic of surprise reactions. Conversely, neutral and sadness states show higher attention weights for long-term features (0.37), indicating that these states are better characterized by sustained patterns rather than immediate responses.

The system's ability to adapt temporal attention based on emotional context is demonstrated through dynamic attention weight changes during emotional transitions. During periods of emotional stability, long-term attention weights increase to capture sustained mood patterns. During emotional transitions, short-term attention weights increase to capture rapid changes in physiological and behavioral indicators.

5.5 Robustness and Generalization Analysis

Robustness evaluation assessed EMOTIONTFN's performance under realistic IoT deployment conditions including environmental variations, sensor failures, and cross-user generalization. The system demonstrated strong robustness across multiple challenging scenarios, confirming its suitability for real-world deployment.

Table 7. Robustness Analysis Results.

Condition	Accuracy Drop (%)	Latency Impact (ms)	Recovery Time (s)
Low Light (50 lux)	2.1	+12	0.8
High Noise (80 dB)	2.8	+8	1.2
User Movement	1.9	+15	0.5
Single Sensor Failure	2.1-4.7	-23	0.3
Two Sensor Failures	6.2-9.1	-45	0.7
Network Latency	0.3	+34	2.1

doi:10.20944/preprints202507.1418.v1

Environmental stress testing included variations in lighting conditions from 50 to 1000 lux, background noise levels from 30 to 80 decibels, and different user mobility patterns. Performance degradation was minimal across all tested conditions, with accuracy drops of less than 3% in most scenarios. The system's robustness to environmental variations is attributed to the multimodal fusion approach and adaptive quality assessment mechanisms.

Sensor failure simulation evaluated the system's ability to handle missing or corrupted modalities through systematic dropout experiments. With single modality failures, accuracy degradation ranged from 2.1% for audio sensor failure to 4.7% for EEG sensor failure. The differential impact reflects the varying importance of different modalities for emotion recognition, with EEG providing unique neural activity information that is difficult to compensate from other sensors.

Even with two simultaneous modality failures, the system maintained over 85% of baseline performance through the adaptive fusion mechanism. The system's graceful degradation capability is crucial for IoT deployment scenarios where sensor reliability may be compromised due to environmental conditions, battery depletion, or hardware failures.

5.6 Real-World Deployment Case Study

To validate EmotionTFN's practical applicability, a comprehensive real-world deployment study was conducted in a smart home environment with 12 participants over a 4-week period. The deployment included wearable sensors integrated into smartwatches for PPG and GSR monitoring, ambient cameras for facial expression analysis, and smart speakers for audio emotion recognition.

Table 8. Real-World Deployment Results

Metric	Week 1	Week 2	Week 3	Week 4	Average
System Uptime (%)	95.2	97.8	98.1	97.6	97.2
Accuracy (%)	91.3	92.1	92.8	92.4	92.2
Avg. Latency (ms)	215	208	203	198	206
User Satisfaction (1-5)	3.8	4.1	4.3	4.2	4.1
Privacy Concerns (1-5)	2.1	1.8	1.6	1.7	1.8

The system successfully operated continuously with 97.2% average uptime over the four-week period. System failures were primarily attributed to network connectivity issues and occasional sensor calibration drift rather than fundamental system malfunctions. The deployment demonstrated the system's reliability for long-term continuous operation in real-world conditions.

Average processing latency in the real deployment was 206ms, slightly higher than laboratory conditions due to network communication overhead and environmental interference. However, the latency remained within acceptable bounds for most emotion-aware applications. User acceptance evaluation showed progressive improvement in satisfaction scores from 3.8 to 4.2 over the deployment period, indicating successful user adaptation and system refinement.

Privacy concerns were minimal and decreased over time, with average scores of 1.8 on a scale where 1 indicates no concerns and 5 indicates major concerns. The low privacy concern ratings are attributed to the system's local processing capabilities, transparent data handling policies, and comprehensive user control mechanisms.

The deployment study revealed several practical considerations for IoT emotion recognition systems. Sensor placement and calibration significantly impact performance, requiring careful attention during installation and periodic recalibration. Environmental factors such as lighting changes and background noise require adaptive algorithms to maintain consistent performance. User behavior patterns vary significantly across individuals, highlighting the importance of personalization and adaptation mechanisms.

6. Discussion

6.1 Technical Contributions and Significance

The Multi-Scale Temporal Fusion Network represents a significant advancement in the field of IoT-based emotion recognition through several key technical contributions. The most important innovation is the multi-scale temporal modeling approach that simultaneously captures emotional dynamics across different time scales. This capability addresses a fundamental limitation of existing emotion recognition systems that typically operate on fixed time windows and fail to capture the complex temporal nature of human emotions.

The hierarchical temporal attention mechanism enables the system to adaptively focus on different time scales based on the current emotional context. This adaptation is crucial for real-world applications where emotional expressions vary significantly in their temporal characteristics. Rapid emotional responses such as surprise or fear require immediate detection through short-term features, while sustained emotional states such as sadness or contentment are better characterized through long-term patterns.

The adaptive fusion mechanism represents another significant contribution, addressing the challenge of integrating multimodal sensor data in dynamic IoT environments. Unlike traditional fusion approaches that use fixed combination strategies, the proposed adaptive fusion learns to weigh different modalities based on their current reliability and relevance to the emotional state. This capability is essential for robust operation in IoT environments where sensor availability and quality may vary due to environmental conditions or hardware limitations.

The comprehensive edge computing optimization approach demonstrates that complex multimodal emotion recognition can be successfully deployed on resource-constrained IoT devices without significant accuracy degradation. The combination of model compression, adaptive sampling, and dynamic resource allocation provides a template for deploying other computationally intensive artificial intelligence applications in IoT environments.

6.2 Implications for IoT Emotion Recognition

The results of this study have far-reaching implications for the development and deployment of emotion recognition systems in IoT environments. The demonstrated ability to achieve high accuracy while maintaining real-time performance on resource-constrained devices opens new possibilities for emotion-aware IoT applications across multiple domains.

In healthcare applications, continuous emotion monitoring can provide valuable insights into patient mental health and treatment effectiveness. The non-intrusive nature of the proposed approach, combined with privacy-preserving processing, addresses key concerns about patient acceptance and data protection. The system's robustness to environmental variations makes it suitable for home-based monitoring scenarios where controlled laboratory conditions are not feasible.

Smart home applications can benefit from emotion-aware automation that adapts to residents' emotional states and preferences. The multi-scale temporal modeling capability enables the system to distinguish between momentary emotional responses and longer-term mood patterns, allowing for appropriate automation responses. Brief frustration might trigger immediate assistance such as adjusting lighting or playing calming music, while persistent sadness could prompt longer-term environmental adjustments or suggestions for social interaction.

Human-computer interaction applications can leverage real-time emotion recognition to create more natural and responsive interfaces. The low latency and high accuracy of EMOTIONTFN enable immediate adaptation of interface behavior based on user emotional state, improving user experience and task performance. Educational applications can adapt content difficulty and presentation style based on student emotional engagement, while entertainment systems can adjust content recommendations based on mood preferences.

6.3 Limitations and Future Directions

While EMOTIONTFN demonstrates significant improvements over existing approaches, several limitations remain that present opportunities for future research. The current system requires initial calibration and training data for optimal performance, which may limit its applicability in scenarios where such data is not readily available. Future work could explore few-shot learning and meta-learning approaches to reduce data requirements for new deployments and enable rapid adaptation to new users or environments.

The evaluation was conducted primarily on established datasets and controlled deployment scenarios. More extensive real-world validation across diverse populations, cultures, and environmental conditions would strengthen the evidence for practical applicability. Long-term longitudinal studies could provide insights into system performance over extended deployment periods and user adaptation patterns.

Cultural variations in emotional expression represent an important consideration for global deployment of emotion recognition systems. The current evaluation included participants from multiple demographic groups, but more extensive validation is needed to ensure fair and accurate performance across different cultural backgrounds and expression patterns. Future research should investigate cultural adaptation mechanisms and develop culturally aware emotion recognition models.

The current emotion model focuses on basic emotions and dimensional representations commonly used in affective computing research. Future extensions could incorporate more sophisticated emotion models that capture complex emotional states such as mixed emotions, emotional transitions, and culturally-specific emotional concepts. Integration with cognitive and social psychology theories could provide deeper insights into emotional processes and improve recognition accuracy.

Privacy and security considerations, while addressed in the current design, require ongoing attention as new threats and regulations emerge. Future work could explore advanced privacy-preserving techniques such as federated learning, homomorphic encryption, and differential privacy to further enhance data protection while maintaining system functionality. The development of standardized privacy frameworks for emotion recognition systems would benefit the entire field.

6.4 Broader Impact and Ethical Considerations

The deployment of emotion recognition systems in IoT environments raises important ethical considerations that must be carefully addressed to ensure responsible development and deployment. Continuous monitoring of emotional states has the potential for both beneficial applications and harmful misuse, requiring clear guidelines and regulations to protect user interests.

Important considerations for future work include addressing potential bias across diverse populations and understanding the psychological impact of continuous emotion monitoring on users. Ethical guidelines and user-centered design principles are essential for responsible deployment.

Data ownership and control represent critical issues for IoT emotion recognition systems. Users must have clear understanding and control over how their emotional data is collected, processed, stored, and used. The development of standardized privacy frameworks and user interfaces for emotional data management is essential for widespread adoption of these technologies.

The potential for misuse of emotion recognition technology by malicious actors or authoritarian governments raises serious concerns about surveillance and social control. Technical safeguards, legal protections, and international cooperation are needed to prevent abuse while enabling beneficial applications of the technology.

7. Conclusions

This paper has presented the Emotion-aware Multi-Scale Temporal Fusion Network (EmotionTFN), a novel architecture for real-time multimodal emotion recognition specifically designed for IoT environments. The proposed system addresses key challenges in IoT-based emotion recognition through innovative multi-scale temporal modeling, adaptive fusion mechanisms, and comprehensive edge computing optimizations.

The experimental evaluation demonstrates significant improvements over existing approaches, with EMOTIONTFN achieving 94.2% accuracy on emotion classification while maintaining processing latency below 200ms on typical IoT hardware. The system shows robust performance across diverse environmental conditions and demonstrates practical applicability through real-world deployment validation. The comprehensive ablation study confirms the importance of each system component and validates the design decisions made in the proposed architecture.

The key contributions of this work include the development of a multi-scale temporal fusion architecture that captures emotional dynamics across different time scales, the implementation of adaptive fusion mechanisms that maintain performance under varying sensor conditions, the design of comprehensive edge computing optimizations that enable real-time processing on resource-constrained devices, and the provision of extensive evaluation demonstrating practical applicability for IoT deployment.

The implications of this research extend beyond technical contributions to broader considerations of ethical deployment, user privacy, and social impact. The demonstrated feasibility of accurate, real-time emotion recognition in IoT environments opens new possibilities for emotion-aware applications while raising important questions about responsible development and deployment practices.

Future research directions include extending the emotion model to capture more complex emotional states, developing privacy-preserving techniques for distributed emotion recognition, investigating cultural adaptation mechanisms for global deployment, and conducting long-term studies to understand the impact of continuous emotion monitoring on user behavior and well-being.

The work presented in this paper represents a significant step toward practical, accurate, and privacy-preserving emotion recognition in IoT environments. The comprehensive evaluation and real-world deployment validation provide evidence for the system's readiness for practical applications, while the identification of limitations and future research directions provides a roadmap for continued advancement in this important field.

The open-source implementation of EmotionTFN will be made available to facilitate reproducible research and accelerate the development of emotion-aware IoT applications. The research community is encouraged to build upon this work and explore new applications and improvements to advance the field of affective computing in IoT environments.

Author Contributions: Conceptualization, S.Y.; Methodology, S.Y.; Validation, S.Y.; Formal Analysis, S.Y.; Writing—Original Draft Preparation, B.K.; Writing—Review & Editing, S.Y.; Funding Acquisition, B.K. The author has read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study used fully de-identified public datasets (MELD, DEAP, and G-REx) that comply with international research ethics standards. According to institutional guidelines and national regulations, IRB approval was waived for secondary analysis of anonymized data.

Informed Consent Statement: Not applicable.

Data Availability Statement: The public datasets used in this study are available as follows: MELD dataset is available at https://affective-meld.github.io/ under academic license; DEAP dataset is available at https://www.eecs.qmul.ac.uk/mmv/datasets/deap/ upon request to the original authors; G-REx dataset is available through the original publication repositories. The real-world deployment data collected during the case study cannot be shared publicly due to privacy regulations and participant consent restrictions. The EmotionTFN implementation code and preprocessing scripts are available at [https://github.com/SungwookYoon/1mstfn] under MIT license and audio overview can check [https://notebooklm.google.com/notebook/5971f1fb-d25e-4dc9-8f35-b1182869b4be/audio]. Researchers interested in accessing specific experimental configurations or additional analysis code may contact the corresponding author with reasonable requests.

Acknowledgments: The authors express sincere gratitude to the 12 participants in the real-world deployment study for their valuable cooperation and feedback throughout the four-week evaluation period. We acknowledge the computational resources provided by the KT Cloud Center and the technical support from the IoT Innovation Laboratory. We thank the Gyeongbuk ICT Cloud Association and the Gyeongbuk Software Industry Association for their deployment research support and collaboration during the IoT device testing phases. Special thanks to the original creators and maintainers of the MELD, DEAP, and G-REx datasets for making their valuable emotion recognition datasets publicly available for research purposes. We also appreciate the anonymous reviewers for their constructive feedback and suggestions that significantly improved the quality of this manuscript, and the technical assistance provided by the Edge Computing Research Group throughout the project.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Atzori, L.; Iera, A.; Morabito, G. The internet of things: A survey. Comput. Netw. **2010**, 54, 2787–2805. https://doi.org/10.1016/j.comnet.2010.05.010
2. Picard, R.W. *Affective Computing*; MIT Press: Cambridge, MA, USA, 1997.
3. Kołakowska, A.; et al. Emotion recognition and its applications. *Hum.-Comput. Syst. Interact.* **2014**, 251–262.
4. Gubbi, J.; Buyya, R.; Marusic, S.; Palaniswami, M. Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Gener. Comput. Syst.* **2013**, *29*, 1645–1660. https://doi.org/10.1016/j.future.2013.01.010
5. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 423–443. https://doi.org/10.1109/TPAMI.2018.2798607
6. Shi, W.; Cao, J.; Zhang, Q.; Li, Y.; Xu, L. Edge computing: Vision and challenges. *IEEE Internet Things J.* **2016**, *3*, 637–646. https://doi.org/10.1109/JIOT.2016.2579198
7. Zeng, Z.; Pantic, M.; Roisman, G.I.; Huang, T.S. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 39–58. https://doi.org/10.1109/TPAMI.2008.52
8. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf. Fusion* **2017**, *37*, 98–125. https://doi.org/10.1016/j.inffus.2017.02.003
9. Lian, H.; Lu, C.; Li, S.; Zhao, Y.; Tang, C.; Yuan, Y. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy* **2023**, *25*, 1661. https://doi.org/10.3390/e25121661
10. Lian, H.; Lu, C.; Li, S.; Zhao, Y.; Tang, C.; Yuan, Y. A survey of deep learning-based multimodal emotion recognition: datasets, methods and challenges. *Appl. Intell.* **2023**, *53*, 9570–9589. https://doi.org/10.1007/s10489-022-04292-9
11. Geetha, A.V.; Darshan, H.; Susrutha, K. Multimodal emotion recognition with deep learning: advancements, challenges, and future directions. *Neurocomputing* **2024**, *573*, 127217. https://doi.org/10.1016/j.neucom.2024.127217
12. Udahemuka, G.; Ruhunage, I.; Kaminduwa Gamage, D.; Priyankara, H.D.N.; Perera, A.S.; Ragel, R. Multimodal emotion recognition using visual, vocal and physiological modalities. *Appl. Sci.* **2024**, *14*, 2155. https://doi.org/10.3390/app14052155
13. Ramaswamy, M.P.A.; Kumar, N.; Venkatesh, S. Multimodal emotion recognition: A comprehensive review. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2024**, *14*, e1503. https://doi.org/10.1002/widm.1503
14. Kim, K.H.; Bang, S.W.; Kim, S.R. Emotion recognition system using short-term monitoring of physiological signals. *Med. Biol. Eng. Comput.* **2004**, *42*, 419–427. https://doi.org/10.1007/BF02344719
15. Shu, L.; Xie, J.; Yang, M.; Li, Z.; Li, Z.; Liao, D.; Yang, X. A review of emotion recognition using physiological signals. *Sensors* **2018**, *18*, 2074. https://doi.org/10.3390/s18072074
16. Ham, S.M.; Choi, Y.J.; Choi, J.W.; Kim, D.H. A negative emotion recognition system with Internet of Things. *Electronics* **2023**, *12*, 1359. https://doi.org/10.3390/electronics12061359
17. Bravo, L.; Villarreal, V.; Cerna, J. A systematic review on artificial intelligence-based multimodal dialogue systems with emotion recognition. *Multimodal Technol. Interact.* **2025**, *9*, 8. https://doi.org/10.3390/mti9010008
18. Sharma, D.; Gupta, A.; Singh, P. Smart emotion detection: An AI and IoT approach to speech analysis. *EELET* **2024**, *3*, 45–52.
19. Lahat, D.; Adali, T.; Jutten, C. Multimodal data fusion: An overview of methods, challenges, and prospects. *Proc. IEEE* **2015**, *103*, 1449–1477. https://doi.org/10.1109/JPROC.2015.2460697

20. Wang, Y.; Guan, L.; Venetsanopoulos, A.N. Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. *IEEE Trans. Multimedia* **2012**, *14*, 597–607. https://doi.org/10.1109/TMM.2012.2186796
21. Chen, M.; Wang, S.; Liang, P.P.; Baltrušaitis, T.; Zadeh, A.; Morency, L.P. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the ACM International Conference on Multimodal Interaction*; ACM: New York, NY, USA, 2017; pp. 163–171. https://doi.org/10.1145/3136755.3136801
22. Bang, J.; Kim, H.; Lee, H. A hybrid multimodal emotion recognition framework for UX evaluation using generalized mixture functions. *Sensors* **2023**, *23*, 3644. https://doi.org/10.3390/s23073644
23. Shi, X.; Zhang, Y.; Wang, L. Multimodal fusion of music theory-inspired and self-supervised representations for improved emotion recognition. In *Proceedings of the Interspeech*; 2024.
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; 2017; Vol. 30.
25. Chen, W.; Xing, X.; Xu, X.; Pang, J.; Du, L. Multimodal emotion recognition based on facial expressions, speech, and EEG. *IEEE Open J. Eng. Med. Biol.* **2023**, *4*, 81–89. https://doi.org/10.1109/OJEMB.2023.3267813
26. Li, J.; Yang, Z.; Zhang, H.; Xu, M.; Zhao, S.; Liu, M.; Sun, M. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2023; pp. 6631–6640.
27. Dai, W.; Cahyawijaya, S.; Liu, Z.; Fung, P. A novel approach for multimodal emotion recognition: Multimodal semantic information fusion. *arXiv preprint* **2024**, arXiv:2407.12173. https://doi.org/10.48550/arXiv.2407.12173
28. Cheng, Z.; Liu, X.; Li, J.; Wang, H.; Zhang, Y. Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*; 2024.
29. Han, S.; Mao, H.; Dally, W.J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint* **2015**, arXiv:1510.00149. https://doi.org/10.48550/arXiv.1510.00149
30. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv preprint* **2015**, arXiv:1503.02531. https://doi.org/10.48550/arXiv.1503.02531
31. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. https://doi.org/10.1007/s10579-008-9076-6
32. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; 2019; pp. 527–536. https://doi.org/10.18653/v1/P19-1050
33. Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.S.; Yazdani, A.; Ebrahimi, T.; Patras, I. DEAP: A database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput.* **2012**, *3*, 18–31. https://doi.org/10.1109/T-AFFC.2011.15
34. Kraack, K. A multimodal emotion recognition system: Integrating facial expressions, body movement, speech, and spoken language. *arXiv preprint* **2024**, arXiv:2406.15063. https://doi.org/10.48550/arXiv.2406.15063
35. Li, J.; Zhang, X.; Huang, L.; Li, F.; Shen, S.; Liu, J.; Shang, L. Multimodal emotion recognition in conversation based on hypergraph. *Electronics* **2023**, *12*, 2664. https://doi.org/10.3390/electronics12122664
36. Lian, H.; Lu, C.; Li, S.; Zhao, Y.; Tang, C.; Yuan, Y. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy* **2023**, *25*, 1661. https://doi.org/10.3390/e25121661

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s)

doi:10.20944/preprints202507.1418.v1

disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.