THIẾT KẾ VÀ TRIỂN KHAI ỨNG DỤNG SỬ DỤNG WIFI CSI ĐỂ CHẨN ĐOÁN CẢM XÚC CON NGƯỜI TRONG CHĂM SÓC SỨC KHỎE TINH THẦN

DESIGN AND IMPLEMENTATION AN APP USING WIFI CSI FOR DIAGNOSING HUMAN EMOTIONS IN MENTAL HEALTHCARE

SVTH: Đặng Minh Quân, Lê Văn Đức, Nguyễn Thị Tuyết Sương, Hoàng Huy Học, Đoàn Nhật Huy

Lớp: 21ECE, 21QLCN2, 22TDT4, 22ES; Khoa: Khoa học công nghệ tiên tiến; Email: 123210100@sv1.dut.udn.vn, 123210094@sv1.dut.udn.vn, 118210190@sv1.dut.udn.vn, 102220189@sv1.dut.udn.vn, 123220054@sv1.dut.udn.vn

GVHD: TS. Nguyễn Quang Như Quỳnh, NCS. Phùng Hữu Tài

Khoa: Khoa học công nghệ tiên tiến, Trường đại học Bách Khoa Đà nẵng, Email: ngnquynh@dut.udn.vn, huutaibkdn@gmail.com

Tóm tắt

Chăm sóc sức khỏe tinh thần đang là một thách thức lớn, đặc biệt tại Việt Nam, nơi có 3,2 triệu người (3,1% dân số) mắc chứng trầm cảm, với tỷ lệ cao nhất (5,4%) ở nhóm thanh niên 18-29 tuổi, đặc biệt là phụ nữ. Các phương pháp theo dõi cảm xúc truyền thống, như tự báo cáo hoặc sử dụng cảm biến đeo trên người, thường xâm lấn, kém chính xác và không phù hợp để đánh giá cảm xúc theo thời gian thực. Chúng tôi đề xuất một phương pháp tiếp cận mới, không xâm lấn, sử dụng Thông tin Kênh WiFi (CSI) để phân tích các thay đổi tinh vi trong nhịp thở nhằm nhận diện cảm xúc. Hệ thống của chúng tôi sử dụng mô hình Vision Transformer (ViT) để phân loại cảm xúc dựa trên biểu diễn quang phổ của tín hiệu CSI. Tuy nhiên, ViT đòi hỏi lượng dữ liệu lớn, gây hạn chế trong quá trình huấn luyện. Để khắc phục điều này, chúng tôi áp dụng Knowledge Distillation, sử dụng mô hình nhận diện cảm xúc qua khuôn mặt làm giáo viên để truyền tri thức cho mô hình CSI. Phương pháp này cải thiện hiệu quả huấn luyện và nâng cao độ chính xác, đạt 72.97% trong việc nhận diện các cảm xúc chính (vui vẻ, tức giận, trung tính, buồn bã).

Từ khóa - WiFi Channel State Information (CSI); Vision Transformer; Nhận diện cảm xúc; Tri thức chưng cất; Chăm sóc sức khỏe tinh thần; Biểu diễn quang phổ; Theo dõi không xâm lấn; Phân tích cảm xúc theo thời gian thực; Học sâu; Nhận diện biểu cảm khuôn mặt.

Abstract

Mental healthcare remains a critical challenge, especially in Vietnam, where 3.2 million people (3.1% of the population) suffer from depression, with the highest prevalence (5.4%) among young adults aged 18-29, particularly women. Traditional emotion monitoring methods, such as self-reports and wearable sensors, are often invasive, unreliable, and unsuitable for real-time assessment. We propose a novel, non-invasive approach using WiFi Channel State Information (CSI) to analyze subtle respiratory patterns for emotion recognition. Our system leverages a Vision Transformer (ViT) model to classify emotions based on spectrogram representations of CSI signals. To address the data-hungry nature of ViTs and improve performance, we apply Knowledge Distillation, using a vision-based facial expression model as a teacher for our CSI-based model. This approach enhances training efficiency and boosts classification accuracy, achieving 72.97% in recognizing key emotions (happy, anger, neutral, sadness).

Key words - WiFi Channel State Information (CSI); Vision Transformer (ViT); Emotion Recognition; Knowledge Distillation; Mental Healthcare; Spectrogram Representation; Non-invasive Monitoring; Real-time Emotion Analysis; Deep Learning; Facial Expression Recognition

1. Introduction

Emotion is an important factor in human social communication. Emotions affect both human physiological and psychological states, playing a very important role in human life. Positive emotions help improve human health and work efficiency, while negative emotions can cause health problems. The long-term accumulation of negative emotions is a factor that can easily lead to depression, which can lead to suicide in the worst cases. Compared with mood, which is a conscious mental state or dominant emotion at a moment, emotion usually refers to a mental state that arises spontaneously rather than through conscious effort, and is often accompanied by physical and physiological changes involving human organs and tissues such as the brain, heart, skin, blood flow, muscles, facial expressions, voice,...

In Vietnam, 3.2 million people (3.1% of the population) suffer from depression, with the highest rate (5.4%) among people aged 18-29, especially women. Due to the complexity of the mutual interaction between physiology and psychology in emotion, the accurate and timely recognition of human emotions is still limited in our understanding and remains the goal of relevant scientific research and industry.

Emotion recognition has been applied in many fields such as safe driving [1], health care especially mental health monitoring [2], social security [3], etc. Emotion recognition methods such as using human physical signals such as facial expressions [4], speech [5], gestures, postures, etc., have the advantage of being easy to collect and have been studied for many years. However, reliability cannot be guaranteed, because it is relatively easy for people to control physical signals such as facial expressions or speech to conceal their true emotions, especially in social communication.

Previous studies, such as [6], have used WiFi CSI to detect emotions through facial movements and breathing,

eliminating the use of image data in the training process. This paper introduces a novel approach to emotion recognition using WiFi Channel State Information (CSI) without relying on traditional image-based methods. We have identified distinct variations in spectrogram representations that correspond to different emotions derived from CSI signals and utilize a Vision Transformer (ViT) model for classification. To overcome the high computational demand of ViT, we pioneer the use of Knowledge Distillation to transfer knowledge from facial expression data to CSI-based emotion recognition, significantly enhancing model performance. Our model achieves a notable accuracy of 72.97%, a 7% improvement over training without Knowledge Distillation, demonstrating the feasibility of CSI-based emotion recognition. Additionally, we explore the impact of different segment lengths (300, 600, and 1200) on performance to determine the optimal segmentation strategy.

2. Related work

Several studies have explored emotion detection using alternative methods. Traditional approaches often rely on image-based techniques, such as facial expression recognition through cameras, to capture visual cues like smiles or frowns for emotion classification. For instance, methods using deep learning models like Convolutional Neural Networks (CNNs) on facial images have achieved high accuracy but raise privacy concerns due to the need for constant visual monitoring. Other works have utilized WiFi signals to detect facial expressions indirectly[2], where WiFi signals were used to infer facial movements, yet these methods still focus on facial features rather than physiological signals. In contrast, our approach leverages WiFi Channel State Information (CSI) to detect emotions through breathing patterns and body movements, offering a non-invasive and privacy-preserving alternative that does not rely on facial data or visual input.

Our contribution: This study makes several key contributions to the field of emotion detection in mental healthcare. We constructed a novel CSI-emotion dataset by capturing breathing patterns using WiFi CSI signals, collected via an ESP32 setup without any image data, ensuring privacy. We pioneered the application of Knowledge Distillation to transfer knowledge from a Vision Transformer (ViT) model for facial expression recognition to a ViT model for CSI-based emotion recognition, improving detection accuracy by approximately 8%. We developed a secure, privacy-preserving emotion detection system that achieves 72% accuracy without relying on cameras or wearable devices, enabling real-time emotion monitoring for applications in education, office, and healthcare settings.

3. Methodology

3.1. Wifi CSI

3.1.1. How Wifi CSI works:

This study utilizes Wi-Fi Channel State Information (CSI) to detect breath rhythm. CSI provides detailed information about the transmission channels between devices in a Wi-Fi network, which can be used to assess channel quality and optimize network parameters for improved data transmission performance [7]. The methodology in this research involves the collection, processing, and analysis of CSI data from transmitted packets to detect variations in the channel characteristics caused by human motion.

CSI captures physical channel measurements at the subcarrier level, providing insights into how transmitted signals interact with the environment. These interactions reveal subtle variations in signal strength due to movement, including slight human body motions such as breathing. As the human body moves—especially during respiration—small changes in the CSI can be detected by analyzing frequency and time-domain features, which in turn allow for the extraction of movement patterns [9]. By examining these variations, we can map CSI data to spatial geometry, enabling real-time monitoring and detection of respiratory rhythms [10].

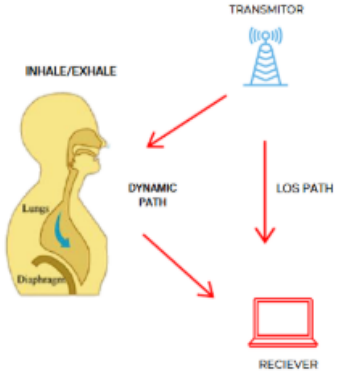


Figure 1: *How wifi CSI interact*

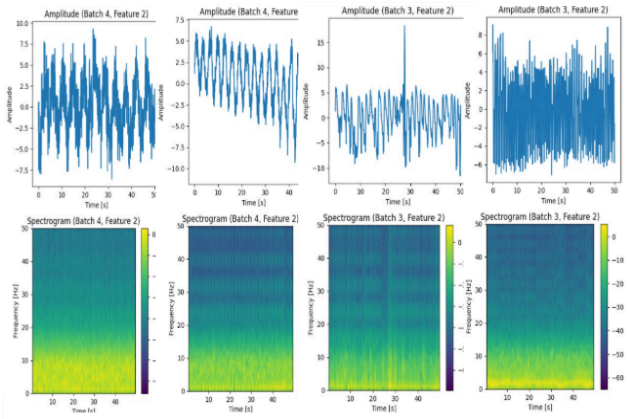
3.1.2 CSI Data Collection

CSI data is gathered through antennas on Wi-Fi devices. In this study, two ESP32 devices (one as a station and one as an access point) are used to collect CSI data. When a data packet is transmitted from the sender to the receiver, information regarding the transmission channel between the two devices is recorded. CSI reflects various channel characteristics, such as signal attenuation, delay, and signal dispersion. These parameters are extracted by analyzing feedback from transmitted signal packets, which are modulated into complex domain indices. CSI is typically derived from the received signal by measuring the amplitude and phase shift of individual subcarriers in the frequency domain, allowing for detailed channel feedback to be captured.

3.1.3 Emotion-Induced Changes in Spectrogram Representations

We discovered that different emotional states influence respiratory patterns, which can be captured through WiFi Channel State Information (CSI). Spectrogram analysis of CSI signals provides a time-frequency representation of these variations, offering valuable insights for emotion recognition. As shown in **Figure 2**, the spectrogram representations of four emotional states—Sadness, Neutral, Happiness, and Anger (from left to right)—demonstrate distinct spectral characteristics. These variations enable a non-invasive approach to monitoring mental states, contributing to advancements in emotion-aware applications for

healthcare and human-computer interaction.



a) Sad b) Neutral c) Happy d) Angry

***Figure 2**: Spectrogram Variations Across Emotions*

These findings highlight the differences between emotions in spectrogram representations of CSI signals. Understanding these variations enables the development of non-invasive emotion recognition systems, paving the way for applications in mental healthcare and human-computer interaction. Each emotion exhibits distinct spectral characteristics:

- Sadness: Strong low-frequency components dominate, indicating slow and deep breathing patterns.
- Neutral: Breathing frequency remains stable and consistent, reflecting a calm physiological state.
- Happiness: High-frequency components are more pronounced, suggesting lighter and faster breathing.
- Anger: Irregular and chaotic breathing rhythms appear, with fluctuations in frequency and intensity.

3.2. Machine Learning Model

The Vision Transformer (ViT) has emerged as a dominant approach in image processing, particularly in facial expression recognition (FER), due to its superior ability to model long-range dependencies through self-attention mechanisms [11]. Unlike traditional convolutional neural networks (CNNs), which rely on localized feature extraction, ViT processes images as sequences of patches, enabling it to capture global contextual relationships more effectively [12]. This architectural advantage has been empirically validated across multiple benchmark datasets, including FER-2013, AffectNet, and RAF-DB, where ViT-based models consistently achieve state-of-the-art performance [13]. For instance, studies have demonstrated that ViT outperforms CNN-based methods by significant margins, particularly in handling occlusions and pose variations, owing to its dynamic attention weighting across facial regions [14]. Furthermore, the scalability of ViT with larger datasets and higher-resolution images further solidifies its position as the leading model for complex visual recognition tasks [15]. Given these advantages, ViT represents not only a theoretical advancement but also a practical solution for advancing facial expression prediction and related computer vision tasks [16].

3.3. Features Knowledge Distillation

Despite the remarkable performance of Vision Transformers (ViTs) in facial expression recognition (FER), their effectiveness heavily relies on large-scale training data, which limits their applicability in scenarios with limited annotated samples [17]. To address this challenge, we employ knowledge distillation (KD), where a compact student ViT model learns from a pre-trained teacher ViT model with identical architecture but trained on a large FER dataset. This approach allows the student model to inherit meaningful features and robust representations from the teacher while reducing dependency on extensive labeled data [18].

Recent studies have demonstrated that KD significantly enhances the efficiency of ViTs by transferring attention-based knowledge, where the student mimics the teacher's attention maps to capture discriminative facial features [19]. Specifically, the teacher model, pre-trained on a comprehensive FER dataset (e.g., AffectNet or RAF-DB), provides attention-guided features that help the student model generalize better even with limited training samples [20]. This is particularly crucial in FER, where subtle expression variations require robust feature representations. Empirical evidence shows that KD-based ViTs achieve comparable accuracy to their teachers while being more computationally efficient [21].

Our approach aligns with findings in [22], where KD between ViTs improved model generalization in data-scarce scenarios. Additionally, [23] highlights that distilling knowledge from attention mechanisms enhances the student's ability to focus on semantically relevant facial regions, such as eyes and mouth, which are critical for expression recognition. By leveraging KD, our method not only mitigates ViT's data hunger but also preserves high accuracy, making it suitable for real-world FER applications with limited labeled data.

4. Experiment Setup:

4.1. ESP32 Configuration:

The hardware setup involves two ESP32 modules. The ESP32-S3 is configured as an Access Point (AP), receiving Wi-Fi signals, while the ESP32-S operates in Station (STA) mode to transmit Wi-Fi packets. This setup enables the collection of Channel State Information (CSI) data, which characterizes the wireless channel as the signal propagates through the environment. The ESP32-S continuously transmits Wi-Fi packets, acting as the data source, while the ESP32-S3, in AP mode, captures the CSI data from these transmitted packets. Both modules are programmed with compatible firmware to ensure smooth communication in a controlled indoor environment. The configuration includes a baud rate of 9600, with UART used for console output and a tick rate of 1000 Hz.

4.2. Environment for Data Collection:

Data is collected in a controlled indoor environment, specifically a sealed room where the participants can sit

still and breathe. The setup ensures that external factors, such as airflow or noise, are minimized, providing consistent and reliable data. Each participant follows a specific breathing pattern associated with each emotion. For anger, participants take deep, irregular breaths with sudden increases in frequency, mimicking the physiological response to heightened emotional arousal. For happiness, participants giggle and let the breath come out of their mouth, reflecting the lighter and more spontaneous nature of happy emotions. For sadness, participants adopt slow, deep breaths, representing the heavy and slow respiratory rhythm commonly associated with feelings of sadness. For neutral, participants maintain a steady, consistent breathing rate to represent a calm, unaltered emotional state. To induce these emotions, participants are shown a series of videos, each designed to evoke a specific emotional response. For the happy emotion, participants watch a joyful video, while for sadness and anger, they view videos that elicit those emotions. A neutral video is shown to induce the neutral emotional state. These videos help standardize the emotional responses and ensure consistent data collection for each emotion category. Additionally, a camera records participants' facial expressions throughout the experiment to capture the visual aspect of their emotional responses, providing further context for emotion analysis.

There are a total of 8 participants in the experiment. all university students, assume an upright posture and breathe through the ESP32 devices as shown in **Figure 3.**

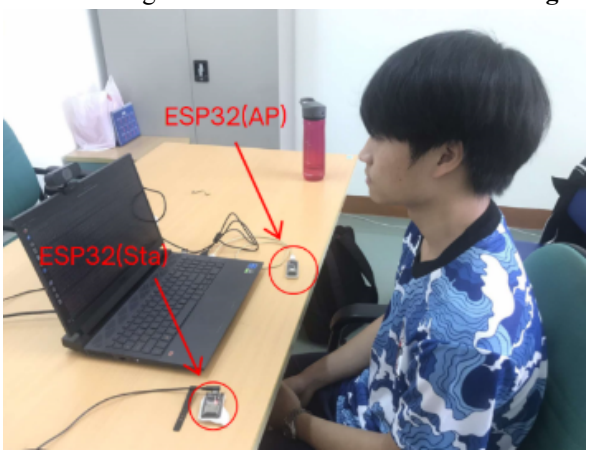


***Figure 3**. setup and participant right posture to collect data*

The data collection process involves capturing WiFi Channel State Information (CSI) at a sampling rate of 100 packets per second. For every 600 consecutive packages (equivalent to 6 seconds of data collection), a spectrogram is generated and paired with a simultaneously captured facial image as 1 sample, ensuring temporal synchronization between wireless signal features and facial expressions. Spectrograms are computed using a sampling frequency (fs) of 10,000 Hz, then resized to 224×224 pixels to match the standard input dimensions of Vision Transformers (ViTs). Corresponding facial images are similarly resized to 224×224 pixels and normalized for model input.

During the session, the ESP32 modules continuously transmit and receive Wi-Fi packets, capturing fluctuations in the Channel State Information (CSI) data as the participants' breathing cycles affect the wireless channel. These fluctuations represent environmental changes induced by the participants' breath. The collected CSI data is processed to extract features that correlate with the breathing rate. The resulting dataset contains synchronized breathing cycles from all participants, which are subsequently used for model training and evaluation. Each data collection session in the experiment lasts for 5 minutes. A total of 1 hour and 30 minutes is recorded for each emotion, resulting in a total of 6 hours of data across all emotions.

4.3. Model Training Setup:

4.3.1. Student and Teacher Model Setup:

We implement a knowledge distillation framework using two ViT models with identical architectures. The teacher model is pretrained on a large-scale facial expression recognition dataset consisting of 48×48 pixel grayscale images across seven emotion categories (angry, disgust, fear, happy, sad, surprise, neutral). When evaluated on our four-class subset (happy, sad, angry, neutral), this teacher model achieves 99% classification accuracy, as demonstrated in the confusion matrix *(Figure 4)*. The student model is initialized with the teacher's pretrained weights, with the first 11 encoder layers (0-10) frozen only the final layers are fine-tuned as in *Figure 5*.

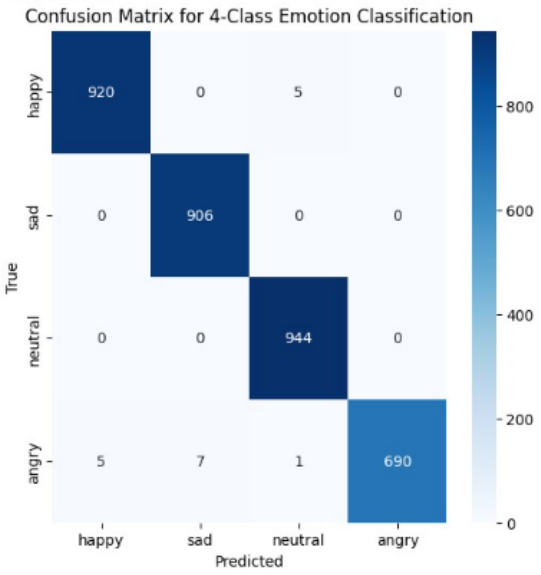


***Figure 4:** Confusion matrix of Teacher model on testset*

4.3.2. Training Loop Setup

Our proposed framework employs a dual-model architecture where paired spectrogram-facial image data serve as distinct inputs to the student and teacher models, as illustrated in *Figure 5*. The spectrograms, generated from WiFi CSI measurements, are processed by the student model, while the corresponding facial images are fed into the pretrained teacher model. We implement feature-level knowledge distillation by aligning the

representations from the last encoder layers of both models. Specifically, the student model is optimized using two key loss components to facilitate effective learning. For feature-level knowledge transfer, we compute the Mean Squared Error (MSE) [24] between the last encoder layer outputs of the student and teacher models, scaled by a factor of 0.001 to ensure balanced optimization, which enables the student model to learn robust intermediate representations from the teacher. The classification objective is enforced through standard cross-entropy loss [25] between the student's final predictions and ground-truth labels. The model is trained end-to-end using the Adam optimizer [26], which combines the benefits of adaptive momentum estimation and RMSprop for efficient parameter updates. This dual-loss approach, combining both high-level feature distillation (via MSE) and label information (via cross-entropy), allows the student model to effectively learn the mapping between WiFi spectrogram patterns and corresponding facial expressions, successfully bridging the modality gap between these two distinct input domains.

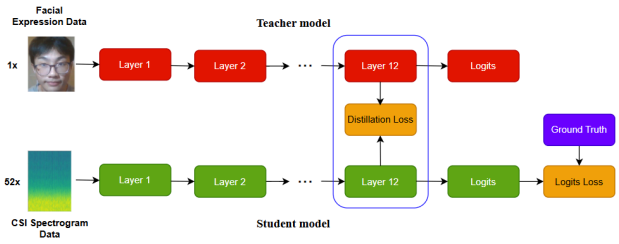


Figure 5: *Feature-Based Knowledge Distillation Diagram*

The dataset was split into three subsets: 70% for training (3500 samples), 15% for validation (700 samples), and 15% for testing (700 samples). This split ensures proper evaluation of the model's performance on unseen data and helps mitigate overfitting.

5. Results:

5.1. Model Performance Comparison:

Our experimental results demonstrate the effectiveness of knowledge distillation in improving model performance. The teacher model, trained exclusively on facial images, achieves a high accuracy of 95.48% on the test set. Through knowledge distillation, the student ViT model processing spectrogram inputs attains an accuracy of 72.97%, representing a significant 7.8% improvement over the baseline ViT trained without distillation (65.18%) as shown in *Figure 6*.

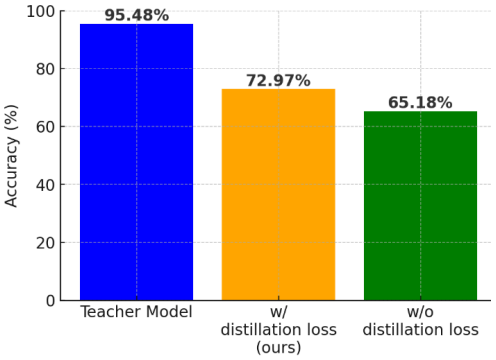


Figure 6: *Comparison of Model predictions on test data*

This substantial performance gain validates our approach of transferring knowledge from the visual domain to the wireless signal domain.

5.2. Model Prediction Accuracy Analysis

The confusion matrix presented in Figure 7 reveals interesting patterns in the model's classification behavior. We observe that the model frequently confuses between sad and neutral emotions, likely due to their similar breathing patterns in the CSI data. Similarly, happy and angry expressions are occasionally misclassified, possibly because they share comparable frequency characteristics in their spectrogram representations.

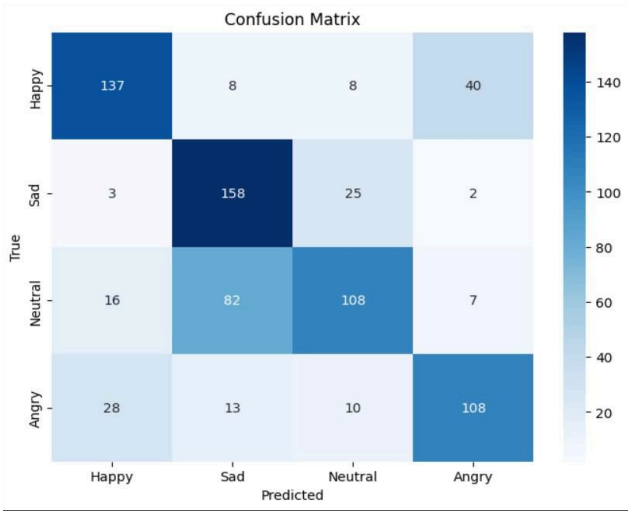


Figure 7: *Confusion matrix of Student model on testset*

As shown in Table 1, the emotion sad achieves particularly strong performance with 70.6% precision and 84% recall. This high predictive accuracy for sadness detection suggests promising applications in mental health monitoring, particularly for depression detection systems. The detailed performance metrics for each emotion category are as follows:

	Precision	**Recall**
Angry	74.5%	71%
Neutral	71.5%	60.7%

Sad	70.5%	84%
Happy	68.8%	69.2%

Table 1: *Precision and Recall of each emotion*

5.3. Impact of different Segment Length

We further investigated how different segment lengths affect model accuracy by testing three configurations are 300, 600 and 1200 packages corresponding to 3, 6 and 12 second. And we gain the accuracy result of each segment length as shown in *Figure 8.*

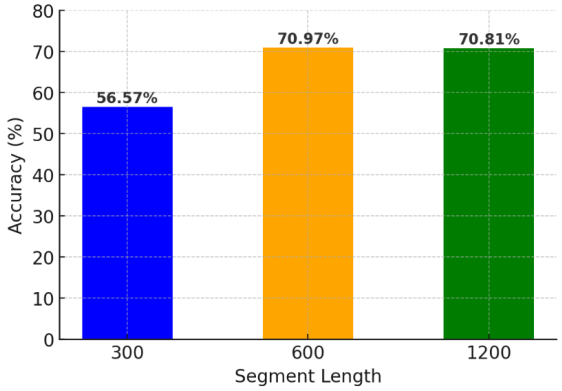


Figure 8: *Accuracy of different segment length of 1200, 600, 300*

The shortest segment length of 300 packets (equivalent to 3 seconds) yields the lowest classification accuracy of 56.57%, as this brief duration proves insufficient for capturing discriminative patterns in spectrogram generation.

In contrast, both 600-packet (6-second) and 1200-packet (12-second) segments demonstrate substantially improved and comparable performance, achieving 72.97% and 70.81% accuracy respectively. This indicates that while sufficient temporal context is crucial (as evidenced by the poor 3-second performance), additional observation time beyond 6 seconds provides diminishing returns in terms of accuracy improvement.

6. Conclusion

This research successfully developed a privacy-preserving emotion detection system using WiFi CSI, achieving a 72.97% accuracy in recognizing emotions through breathing patterns and body movements, without relying on image-based data. By pioneering the use of Knowledge Distillation to transfer knowledge from a Vision Transformer (ViT) model for facial expression recognition to a CSI-based ViT model, we improved detection accuracy by 7.8% compared to a baseline without distillation. Our novel CSI-emotion dataset and optimal segment length analysis (600 packets, 6 seconds) further enhance the system's applicability, demonstrating its potential for real-time emotion monitoring in mental healthcare, education, and workplace settings.

Tài liệu tham khảo

[1] S. De Nadai, M. D'Incà, F. Parodi, M. Benza, A. Trotta, E. Zero, L. Zero, and R. Sacile, "Enhancing safety of transport by road by on-line monitoring of driver emotions," in *Proceedings of the 2016 11th System of Systems Engineering Conference (SoSE)*, Kongsberg, Norway, 12–16 June 2016, pp. 1–4.

[2] R. Guo, S. Li, L. He, W. Gao, H. Qi, and G. Owens, "Pervasive and unobtrusive emotion sensing for human mental health," in *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare*, Venice, Italy, 5–8 May 2013, pp. 436–439.

[3] B. Verschuere, G. Crombez, E. Koster, and K. Uzieblo, "Psychopathy and Physiological Detection of Concealed Information: A review," *Psychol. Belg.*, vol. 46, pp. 99–116, 2006.

[4] Y. D. Zhang, Z. J. Yang, H. M. Lu, X. X. Zhou, P. Phillips, Q. M. Liu, and S. H. Wang, "Facial Emotion Recognition based on Biorthogonal Wavelet Entropy, Fuzzy Support Vector Machine, and Stratified Cross Validation," *IEEE Access*, vol. 4, pp. 8375–8385, 2016.

[5] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks," *IEEE Trans. Multimed.*, vol. 16, pp. 2203–2213, 2014.

[6] Y. Chen, R. Ou, Z. Li, and K. Wu, "WiFace: Facial Expression Recognition Using Wi-Fi Signals," *IEEE Trans. Mobile Comput.*, vol. 21, no. 1, pp. 378–391, Jan. 2022, doi: 10.1109/TMC.2020.3001989.

[7] X. Zhang, L. Zhang, Z. Zhang, et al., "Channel State Information (CSI)-Based Wireless Localization," *IEEE Access*, vol. 6, pp. 73023–73033, 2018.

[8] H. Xu, Y. Chen, J. Liu, et al., "CSI-based Localization and Tracking in Wi-Fi Networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 11, pp. 2451–2463, Nov. 2014.

[9] C. Wu, F. Zhang, Y. Hu, and K. J. R. Liu, "GaitWay: Monitoring and Recognizing Gait Speed Through the Walls," *IEEE Trans. Mobile Comput.*, 2020.

[10] F. Zhang, C. Chen, B. Wang, and K. J. R. Liu, "WiSpeed: A Statistical Electromagnetic Approach for Device-Free Indoor Speed Estimation," *IEEE Internet of Things J.*, 2018.

[11] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv:2010.11929, 2020.

[12] S. Khan et al., "Transformers in Vision: A Survey," ACM Computing Surveys, vol. 54, no. 10s, 2022.

[13] Z. Zhao et al., "Facial Expression Recognition Using Vision Transformers with Cross-Dataset Evaluation," IEEE Transactions on Affective Computing, 2022.

[14] K. Wang et al., "Transformer-Based Facial Expression Recognition: A Comparative Study on FER-2013 and AffectNet," Proc. CVPR Workshops, 2021.

[15] H. Touvron et al., "Training Data-Efficient Image Transformers & Distillation Through Attention," Proc. ICML, 2021.

[16] Y. Li et al., "Vision Transformer for Facial Expression Recognition in the Wild," Proc. ICCV, 2023.

[17] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR*, 2021.

[18] G. Hinton et al., "Distilling the Knowledge in a Neural Network," *arXiv:1503.02531*, 2015.

[19] S. Chen et al., "Cross-Architecture Knowledge Distillation for Vision Transformers," *CVPR*, 2022.

[20] Z. Zhao et al., "Knowledge Distillation for Facial Expression Recognition with Vision Transformers," *IEEE T-PAMI*, 2023.

[21] J. Ba and R. Caruana, "Do Deep Nets Really Need to be Deep?," *NeurIPS*, 2014.

[22] Y. Tian et al., "Contrastive Representation Distillation," *ICLR*, 2020.

[23] L. Yuan et al., "Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet," *ICCV*, 2021.

[24] C. M. Bishop, "Pattern Recognition and Machine Learning," Springer, 2006.
[25] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning," MIT Press, 2016.
[26] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," ICLR, 2015.