Preprint Version

This is the preprint version of the article. Copyright belongs to IEEE. The published version is available online at IEEE Xplore: H. A. Imran, Q. Riaz, K. Hamza, S. Muhammad, and B. Krüger, "From Steps to Sentiments: Cross-Domain Transfer Learning for Activity-Based Emotion Detection in Wearable IoT Systems," *IEEE Internet of Things Journal*, 2026, doi: 10.1109/JIOT.2026.3666469.

Please cite the published version.

From Steps to Sentiments: Cross-Domain Transfer Learning for Activity-Based Emotion Detection in Wearable IoT Systems

Hamza Ali Imran, Qaiser Riaz*, *Senior Member, IEEE*, Kiran Hamza, Shaida Muhammad, Björn Krüger

***Abstract*—Context-aware, gait-based sentiment analysis and emotion perception is an emerging research area within Internet of Things (IoT), aiming to make smart systems more intuitive and responsive. Recognizing emotions from wearable inertial sensor data is challenging due to subtle and compound emotional cues, variability across individuals and contexts, and limited, imbalanced datasets. To address these challenges, we propose Jazbat-Net, a lightweight neural network that leverages Transfer Learning (TL). The model is first trained on a large-scale, publicly available multi-activity dataset collected using wearable inertial sensors, and then retrained on a multi-class emotion dataset, effectively transferring knowledge from the pretraining phase. We evaluate Jazbat-Net with and without TL, across both smartwatch and smartphone based data, and for input dimensions ranging from 1D to 6D. The best results are achieved when pretrained on smartphone-based activity data and retrained on smartphone-based emotion data using a 1D input size. The proposed model attains an average classification accuracy of 95%, with a precision score of 95%, a recall score of 97%, and an F1-score of 96%. Moreover, Jazbat-Net achieves a low theoretical time complexity and requires only $\approx$ 6.96 M Multiply–Accumulate Operations (MACs), which is about 95% fewer computations than the previous State-of-the-Art (SOTA) model. Its space complexity is also low, with a model size of only $\approx$ 110 KB and peak activation memory of $\approx$ 0.35 MB. On-device evaluation on a Xiaomi 13T smartphone demonstrates that Jazbat-Net achieves a median inference latency of only $\approx$ 90.96 ms with a TFLite 32-bit floating point precision (FP32) model size of just $\approx$ 0.158 MB, making it $\approx$ 20$\times$ smaller and $\approx$ 20% faster than the previous SOTA model while maintaining comparable accuracy.**

***Index Terms*—human emotions recognition, transfer learning, gait analysis, human activity recognition, wearable sensors, inertial sensors, Internet of Things, IoT for smart healthcare**

I. Introduction

Overtime, the Internet of Things (IoT) has enabled a wide range of smart applications, including smart homes with automated lighting, energy management, and security systems [1] as well as healthcare solutions such as remote patient monitoring, fall detection in elderly care, medication adherence tracking, and chronic disease management through wearable sensors [2]. Modern IoT systems go beyond simple data collection; for example, smart homes now integrate environmental information to automatically adjust devices such as air conditioners and lighting to enhance user comfort [3]. These advances highlight the importance of creating seamless interaction between humans and IoT devices. Within this context, context-aware, gait-based sentiment analysis has emerged as a critical research direction. By analyzing gait patterns captured through wearable inertial sensors, this field explores how human movement reflects emotional states, personality traits, and health conditions [4].

Human gait analysis has demonstrated its potential for several applications, including assessing motor functions in patients with Parkinson's Disease (PD) [5], identifying fall risks, and detecting gait deviations in nonclinical settings. Its applications extend to medicine, sports, and ergonomics, showing promising outcomes [6], [7]. For monitoring and quantifying motor symptoms in PD, wearable Inertial Measurement Units (IMUs) have been used to track motor fluctuations, dyskinesia, tremors, bradykinesia, freezing of gait, and gait disturbances, aiding patient care and research [8]. Moreover, Gait analysis also holds significant value as it reflects traits like emotional states [9] and soft biometrics such as age and gender. The uniqueness of each individual's walking style [10] further establishes gait as a robust and reliable identifier.

Human Activity Recognition (HAR) using wearable IMUs is a prominent area of research, particularly in the domains of smart devices, healthcare, and fitness applications [11]. Wearable IMUs are employed to record Activities of Daily Living (ADL), playing a pivotal role in health monitoring, disability assistance, and elderly care. Leveraging machine learning, the collected data is analyzed and utilized to classify ADL, thereby promoting the health and well-being of individuals. However, machine learning-based HAR methods often face challenges such as high computational costs, substantial data requirements, and limited scalability across diverse environments.

Emotions play a significant role in shaping actions, decisions, and relationships, influencing social interactions and fostering bonds. Expressed through both verbal and nonverbal cues during activities of daily living, emotions profoundly impact personal and collective experiences. Defined as complex mental states, emotions encompass subjective experiences, physiological responses, and behavioral expressions [12], [13], with reactions varying significantly across individuals. Human Emotion Recognition (HER) has diverse applications, including autism therapy, intelligent chatbots, and immersive technologies. These advancements hold the potential to revolutionize human-technology interactions by enabling systems that are intuitive, empathetic, and responsive to emotions [14].

Hamza Ali Imran, Qaiser Riaz, Kiran Hamza and Shaida Muhammad are with the Department of Computing, School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan. (e-mails: {himran.mscs18seecs, qaiser.riaz, kmehmood.mscs19seecs, smuhammad.mscs19seecs}@seecs.edu.pk)
Björn Krüger is heading the group for Personalized Digital Health and Telemedicine at the Department for Epileptology, University Hospital Bonn, 53127 Bonn, Germany (e-mail: bkrueger@uni-bonn.de). (Corresponding author: Qaiser Riaz; qaiser.riaz@seecs.edu.pk)

Recent research highlights the growing significance of wearable inertial sensors in human sentiment analysis. Embedded in devices such as smartwatches and fitness trackers, these sensors capture motion data associated with physical activities like walking, running, and gesturing [15]. By analyzing gait patterns, posture, and movement dynamics, systems can infer emotional states such as happiness, anger, and sadness. Studies demonstrate that features like step variability, stride length, and acceleration magnitude are closely correlated with emotional responses.

Many existing human motion analysis approaches using wearable IMUs struggle to generalize, limiting real-world use. Although effective in controlled datasets, these models often fail in diverse environments, populations, and sensor setups due to reliance on data-specific features and rigid model designs [16]. Even slight changes in sensor placement or hardware can degrade performance [17]. A key challenge is handling cross-domain shifts, such as demographic variation or transitions from laboratory to real-world conditions [18]. While some works explore domain adaptation, such methods remain non-standard and often require complex retraining [19]. To address these issues, we adopt a two-phase TL strategy to improve generalization in activity and emotion recognition. First, a deep model is pre-trained on a large, public, and multi-activity HAR dataset to learn broad gait patterns across varied contexts. Then, it is fine-tuned on a smaller dataset for emotion recognition, aligning general features with emotion-specific patterns. This design addresses data scarcity and domain shifts, thereby enhancing the model's robustness across users and settings [20].

We introduce Jazbat-Net, a novel deep learning model that combines Bidirectional Gated Recurrent Units (BiGRU) and Multi-Kernel Convolutional Neural Network (MultiCNN) layers. Lightweight yet effective, Jazbat-Net captures generalized temporal dependencies critical for gait-based emotion analysis. The key features of the model and the study contributions are outlined below:

1) "Jazbat-Net" a (BiGRU-MultiCNN) model of approximately 29K trainable parameters trained with the virtue of TL, first trained on a large-scale ADL dataset of smartwatch and smartphone IMUs to learn general human motions related features. These features are transferred by retraining on the small SEECS Emotions dataset (six emotions). The model achieved 95.36% accuracy (Section IV-A1).
2) A systematic evaluation across sensor configurations compared models with and without TL. Configurations included 6D smartwatch $(a_x^w, a_y^w, a_z^w, \omega_x^w, \omega_y^w, \omega_z^w)$, 6D smartphone $(a_x^p, a_y^p, a_z^p, \omega_x^p, \omega_y^p, \omega_z^p)$, 3D accelerations from both devices, and 1D acceleration magnitudes $(\widehat{mag_a^w}, \widehat{mag_a^p})$. Smartphone acceleration magnitude $(\widehat{mag_a^p})$ pretraining achieved the highest accuracy of 95.36%, a 10.26% gain over the baseline without TL (Section IV).
3) The model is invariant to hardware and sampling rate. It was pre-trained on HAR data from Nexus 5/5X, Galaxy S5, and a smartwatch at (20 Hz), and fine-tuned on emotion data from Galaxy S5 at (75 Hz). Despite no down-sampling, it performed well across heterogeneous devices and rates (Section III-A).
4) Time Complexity and Space Complexity Analysis of the model has been performed and compared with SOTA (Section IV-E). Moreover, inference has been done on an Android Phone for real-world inference analysis (Section IV-C).

II. Related Work

HAR using inertial sensors has emerged as a mature and data-rich field. Additionally, the origin of data is of a close nature to gait-based emotions recognition, referred to as HER, making it a suitable pretext task for TL in downstream applications. Although ample work exists within HAR knowledge transfer to enhance generalization across users, devices, and domains, but cross-domain TL is quite limited. Pandit et al. [21] transformed 3D IMUs signals into grayscale images for classification using Inception-v3. Although these methods achieve high accuracy, their reliance on multimodal or image-transformed data limits their suitability for real-time, low-power wearable devices. These studies show the effectiveness of deep learning in HAR and justify its use as a base task for developing transferable representations. Pei et al. [22] introduced novel technique, which blends synthetic and real IMUs data for robust recognition. Kuo et al. [23] demonstrated high accuracy in gesture and pose recognition for human–robot interaction using TL. These studies also validate the utility of pretraining on HAR for feature extraction, but they focus exclusively on activity or interaction tasks.

Fu et al. [24] used pseudo-labeling for personalized HAR, and Link et al. [25] transferred knowledge from volleyball to Ultimate Frisbee throw classification. Celik et al. [26] and Zhang et al. [27] applied image-based CNNs and propagation-based techniques to clinical and sports activity data. While these approaches are effective, they rely on task-specific models or preprocessing pipelines that are not easily generalizable. Domain-adversarial learning [28] aligns features across domains (e.g., users/devices) under closed-set Unsupervised Domain Adaptation, where source and target share the same label space. While effective for handling cross-user shift within HAR [29], it is not directly applicable to our setting, which involves cross-task transfer from HAR to HER with distinct label spaces and objectives. Adversarial methods assume shared classes (e.g., "walk" vs. "run"), whereas we transfer generic gait features to emotion recognition, which lacks such labels. A lightweight adversarial head could still be added within HER to reduce subject identity cues, but this addresses a different problem than our HAR→HER transfer.

In contrast to the abundance of HAR research, emotion recognition using inertial data HER remains comparatively underexplored. Imran et al. [9] proposed a deep neural network achieving 95.23% accuracy in emotion classification using the same dataset we have used in this study, but the model developed and presented was 700K plus trainable parameters, making it a highly complex model. This work is the previous SOTA, and a detailed complexity analysis has been done in

Section (IV-E). Hamza et al. [10] introduced Generisch-Net, a BiGRU-CNN model which they trained separately on HAR, person re-identification, and HER datasets. This study also used the same emotions dataset as ours and achieved only 78.20% accuracy.

These studies confirm that HER using only IMUs data is both feasible and promising, particularly for wearable and resource-constrained applications. However, a unified framework that leverages HAR as a source domain to transfer generalizable features for HER remains underdeveloped, and this idea has merit since data for HAR is available in abundance. The presented study addresses this gap by proposing a lightweight, transferable model that bridges HAR and HER domains.

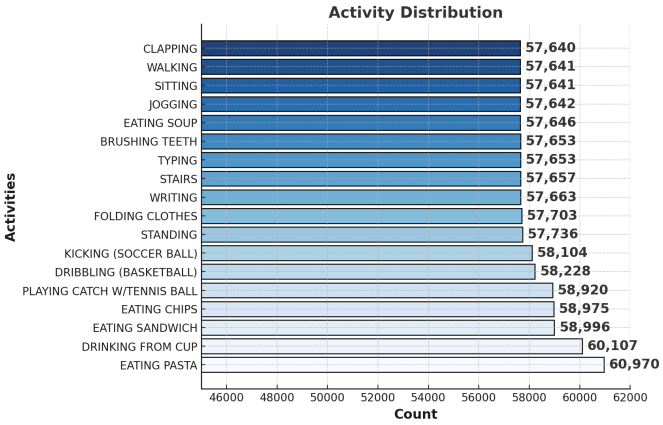


Fig. 1: Activity distribution in the WISDM 2019 dataset, showing the number of segmented samples per activity. The dataset is nearly balanced, with most classes having approximately 57,000 to 61,000 segments. Note that the x-axis does not start at zero, and the unit is segment count, which should be considered when interpreting class proportions, where each segment consists of 256 time steps sampled at 20 Hz.

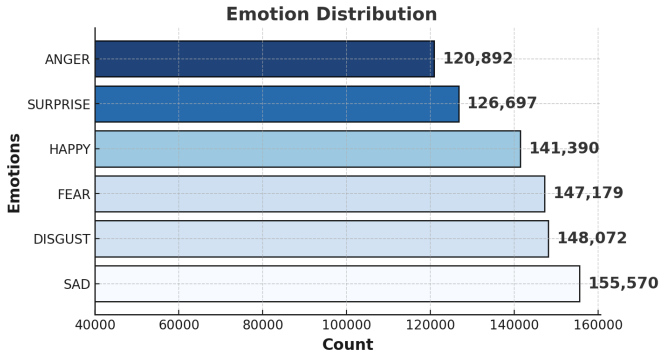


Fig. 2: Class distribution of the SEECS Emotions dataset, showing the number of segmented samples per emotion. The dataset exhibits moderate class imbalance, with SAD having the highest and ANGER the lowest number of segments. Each segment consists of 256 time steps recorded at a sampling rate of 75 Hz. Note that the x-axis does not start at zero, and the unit is segment count, which should be considered when interpreting class proportions.

III. Methodology

This section outlines the proposed methodology. It begins with a description of the datasets, followed by the preprocessing steps, model architecture, and the TL strategy. The complete computational workflow is summarized in Algorithm 1.

A. Datasets

This subsection provides details about the two datasets used in the study. The first is a publicly available HAR dataset, while the second is a locally collected closed access named as SEECS Emotions dataset. The former dataset is utilized for pretraining, whereas the latter is employed for transferring knowledge.

1) Dataset for Pretraining: We used the publicly available WISDM 2019 dataset [30] for model pretraining. It comprises inertial recordings of 51 volunteers aged 18 and 25 years (mean age 21.5, SD not reported). Each participant performed 18 different activities of daily living while inertial data was captured using an LG G Watch worn on the wrist and a smartphone (Google Nexus 5/5X or Samsung Galaxy S5) placed in the participant's pocket. Both devices recorded tri-axial accelerometer and gyroscope signals at a sampling rate of 20 Hz. Following our standard procedure [9], we segmented the data into windows of 256 time steps (12.8 seconds) with a stepping size of 32 samples. These parameters were selected to preserve temporal continuity while ensuring segment independence. The resulting segments were labeled according to activity classes and used for model pretraining. The class distribution is shown in Figure 1, which is nearly balanced across all 18 activities.

2) Dataset for Transferring Knowledge: For TL, we used the SEECS Emotions dataset [31], a closed-access corpus collected from 40 healthy participants (26 males and 14 females), with an average age of 25.2 ± 5.9 years and an average height of 171.6 ± 8.4 cm. A Samsung Galaxy S5 smartphone, worn on the chest and equipped with an MPU-6500 inertial sensor, recorded tri-axial accelerometer and gyroscope data at a sampling rate of 75 Hz. The participants walked naturally on a 40-meter path while recalling personal events corresponding to six basic emotions defined by Ekman and Friesen [32]: *happy*, *fear*, *sad*, *disgust*, *anger*, and *surprise*. Following the same segmentation strategy as in the pretraining phase of study [9], we segmented the data into fixed length windows of 256 time steps (3.4 seconds), with a stepping size of 16 samples. These settings were optimized to capture gait dynamics related to emotional states while maintaining a consistent input structure. The emotion-wise class distribution, shown in Figure 2, is moderately imbalanced.

B. Preprocessing

One of the key objectives of the proposed work is to enhance computational efficiency, which can be achieved by reducing the input dimensionality. Let $\mathbf{X}$ represent the raw input signal of 3D accelerations, denoted as:

$$\mathbf{X} \in \mathbb{R}^{256\times 3}$$

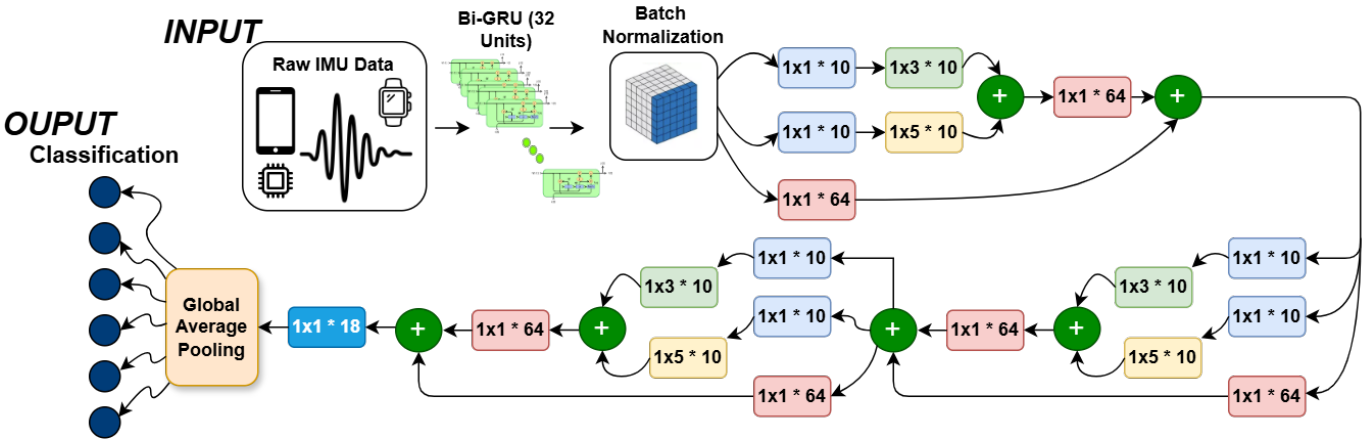


Fig. 3: The architecture of the proposed model.

Algorithm 1 Pretraining Jazbat-Net on HAR Data and TL the Model to Human Emotion Recognition Data

User Input: Select Data Dimensions:
  1D: $\widehat{\text{mag}}_a$ or **3D**: (a_x, a_y, a_z) or
  6D: $(a_x, a_y, a_z, \omega_x, \omega_y, \omega_z)$
  Set segment_dim $= 1, 3,$ or 6 accordingly
Preprocess HAR Data:
  $D_{\text{HAR}} \leftarrow$ Select Data Dimensions
  $D_{\text{seg}} \leftarrow$ Segment(D_{HAR}, segment_dim, 256)
Initialize and Train Jazbat-Net:
  $Model_{\text{Jaz}} \leftarrow$ Initialize with Random Parameters
  $Model_{\text{Jaz}} \leftarrow$ Train($Model_{\text{Jaz}}, D_{\text{seg}}$, 40 epochs)
Modify Jazbat-Net for Emotion Detection:
  $Model_{\text{updated}} \leftarrow$ Remove Last Layer($Model_{\text{Jaz}}$)
  $Model_{\text{updated}} \leftarrow$ Add Softmax Layer (6 nodes)
Preprocess Emotion Data:
  $D_{\text{Emotion}} \leftarrow$ Data Dimensions Selected for HAR
  $D_{\text{Emotion\_seg}} \leftarrow$ Segment(D_{Emotion}, segment_dim, 256)
Train the Updated Model on Emotion Data:
  $Model_{\text{updated}} \leftarrow$ Train($Model_{\text{updated}}, D_{\text{Emotion\_seg}}$, 90 epochs)
Evaluate the Model:
for each X_{batch} in D_{Test} **do**
    $Y_{\text{Emotion}} \leftarrow$ Predict($Model_{\text{updated}}, X_{\text{batch}}$)
    $(F1_{\text{batch}}, Recall_{\text{batch}}, Precision_{\text{batch}}) \leftarrow$ Calculate Metrics($Y_{\text{Emotion}}, Y_{\text{True}}$)
    $F1 \leftarrow F1 + F1_{\text{batch}}$, $Recall \leftarrow Recall + Recall_{\text{batch}}$
    $Precision \leftarrow Precision + Precision_{\text{batch}}$, $N \leftarrow N + 1$
end for
$F1 \leftarrow F1/N$, $Recall \leftarrow Recall/N$, $Precision \leftarrow Precision/N$
Output:
$Output \leftarrow$ {F1-Score : $F1$, Recall : $Recall$, Precision : $Precision$}
return $Output, Model_{\text{updated}}$

where T=256 represents the segment size and the three dimensions correspond to the acceleration components a_x, a_y, and a_z. The acceleration components can be transformed into a 1D representation, $\widehat{\text{mag}}_a \in \mathbb{R}^{256}$, by computing the magnitude of the 3D accelerations vector. The magnitude, $\widehat{\text{mag}}_a$, is defined as:

$$\widehat{\text{mag}}_a[t] = \sqrt{a_x[t]^2 + a_y[t]^2 + a_z[t]^2} - g, \quad \forall t \in \{1, 2, \ldots, S_{\text{Wsize}}\} \tag{1}$$

where $a_x[t]$, $a_y[t]$, and $a_z[t]$ are the components of 3D accelerations along the x-, y-, and z-axes at time step t, and $g = 9.81\,\text{m/s}^2$ is the gravitational constant. The segmentation window size is hence $S_{\text{Wsize}} = 256$, and the stepping size is $S_A = 32$ for the Activities dataset and $S_E = 16$ for the Emotions dataset. The transformation reduces the dimensionality of the input from $\mathbf{X} \in \mathbb{R}^{S_{\text{Wsize}} \times 3}$ to $\widehat{\text{mag}}_a \in \mathbb{R}^{S_{\text{Wsize}} \times 1}$, simplifying the data representation for the neural network. This pre-processing step offers several advantages.

a) Orientation Invariance: The magnitude $\widehat{\text{mag}}_a[t]$ is invariant to the sensor's orientation, as it depends only on the norm of the acceleration vector rather than the specific directions of a_x, a_y, and a_z. This ensures robustness in scenarios where the sensor's orientation varies, such as when a smartwatch is worn on a moving wrist.

b) Feature Robustness: The neural network learns features that are less dependent on specific sensor orientations, focusing instead on consistent motion patterns across different orientations. This enhances the model's reliability and effectiveness in diverse deployment conditions.

c) Dimensionality Reduction: Transforming the input from $\mathbb{R}^{256\times3}$ to $\mathbb{R}^{256\times1}$ reduces computational complexity during training and inference, making the approach more suitable for real-time systems.

After segmentation, each input segment, T, consists of 256 time steps and either 3 channels (for raw acceleration) or 1 channel (for magnitude-transformed acceleration). Thus, the input dimensionality is $\mathbb{R}^{256\times3}$ for raw 3D signals and $\mathbb{R}^{256\times1}$ for orientation-invariant magnitude features. These dimensionalities define the shape of the input passed to the neural network, where each batch contains multiple such

segments. No magnitude transformation was applied to the gyroscope data, in accordance with previous studies [9], [33], which showed minimal performance gain for the magnitudes of the angular velocity.

We segmented each IMUs stream with a fixed-length sliding window and tuned the segmentation hyperparameters via a grid search over window length $S_{\text{Wsize}} \in \{64, 128, 256, 512\}$ and step size $S_{\text{step}} \in \{16, 32, 64, 128\}$. Tuning used 5-fold cross-validation on the training set, optimizing validation accuracy. For the emotion-recognition experiments, the best setting was $S_{\text{Wsize}} = 256$ and $S_E = 16$ (overlap $\approx 93.75\%$), which we use in all reported emotion results. For the activities dataset, we use $S_{\text{Wsize}} = 256$ and $S_A = 32$ (overlap $\approx 87.5\%$), consistent with our previous study as well [9].

C. Architecture

The architecture of the proposed model is illustrated in Figure 3. The model processes raw IMUs data captured from human gait. Depending on the use case, the input sequence can be:

- 1D: The magnitude of 3D accelerations which is denoted as $\widehat{mag_a}$. To distinguish between smartwatch and smartphone data, we use $\widehat{mag_a^w}$, where *w* represents the watch, and $\widehat{mag_a^p}$, where *p* represents the smartphone.
- 3D: The raw accelerations in three axes which are denoted as a_x, a_y, a_z. To distinguish between smartwatch and smartphone data, we use a_x^w, a_y^w, a_z^w, where *w* represents the watch, and a_x^p, a_y^p, a_z^p, where *p* represents the smartphone.
- 6D: A combination of 3D accelerations and 3D angular velocities which are denoted as $a_x, a_y, a_z, \omega_x, \omega_y, \omega_z$. To distinguish between smartwatch and smartphone data, we use $a_x^w, a_y^w, a_z^w, \omega_x^w, \omega_y^w, \omega_z^w$, where *w* represents the watch, and $a_x^p, a_y^p, a_z^p, \omega_x^p, \omega_y^p, \omega_z^p$, where *p* represents the smartphone.

The input, $\mathbf{X} \in \mathbb{R}^{S_{\text{Wsize}} \times F}$, where $S_{\text{Wsize}} = 256$ represents the number of timesteps in each input segment and F is the number of input features, is first processed through a Bidirectional GRU layer with 32 units. This layer captures temporal dependencies, producing an output representation $\mathbf{H}_{\text{GRU}} \in \mathbb{R}^{S_{\text{Wsize}} \times 64}$. To enhance stability during training, batch normalization is applied, resulting in $\mathbf{H}_{\text{BN}}$.

After the BiGRU layer three convolutional modules are applied. These modules are the core of the model consists of three Multi-Scale Convolutional Paths, each designed to extract diverse spatial features. In each module, the input feature map $\mathbf{R}_{i-1}$ (where $\mathbf{R}_0 = \mathbf{H}_{\text{BN}}$) is processed through two parallel branches:

- Branch 1: A 1×1 convolution with 10 kernels, followed by a 1×5 convolution with 10 kernels.

$$\mathbf{B}_{1,i} = \text{Conv}_{1\times1}^{10}(\mathbf{R}_{i-1}) \rightarrow \text{Conv}_{1\times5}^{10}(\mathbf{B}_{1,i}).$$

- Branch 2: A 1×1 convolution with 10 kernels, followed by a 1×3 convolution with 10 kernels.

$$\mathbf{B}_{2,i} = \text{Conv}_{1\times1}^{10}(\mathbf{R}_{i-1}) \rightarrow \text{Conv}_{1\times3}^{10}(\mathbf{B}_{2,i})$$

The outputs of both branches are combined through element-wise summation, generating a unified feature map that captures diverse spatial and temporal patterns.

$$\mathbf{M}_i = \mathbf{B}_{1,i} + \mathbf{B}_{2,i}$$

This combined feature map is then aggregated using a 1×1 convolution with 64 kernels.

$$\mathbf{M}'_i = \text{Conv}_{1\times1}^{64}(\mathbf{M}_i)$$

To incorporate residual learning, a 1×1 convolution with 64 kernels is applied to the module input, and the result is added to the aggregated features.

$$\mathbf{R}_i = \mathbf{M}'_i + \text{Conv}_{1\times1}^{64}(\mathbf{R}_{i-1})$$

where $i = 1, 2, 3$ for the three modules.

After the 3^{rd} module, the output $\mathbf{R}_3$ is passed through a 1×1 convolution with 18 kernels to reduce the feature map dimensionality.

$$\mathbf{F}_{\text{reduce}} = \text{Conv}_{1\times1}^{18}(\mathbf{R}_3)$$

Global average pooling is then applied to $\mathbf{F}_{\text{reduce}}$, which reduces each feature channel to its mean value across the temporal dimension.

$$z_k = \frac{1}{T}\sum_{t=1}^{T} F_{\text{reduce},t,k},$$

where z_k is the pooled value for the k-th feature channel, $F_{\text{reduce},t,k}$ represents the value at timestep t in the k-th channel of $\mathbf{F}_{\text{reduce}}$, and T is the number of timesteps. This operation produces a compact feature vector:

$$\mathbf{z} = [z_1, z_2, \ldots, z_{18}] \in \mathbb{R}^{18}.$$

Finally, the feature vector $\mathbf{z}$ is passed through a dense layer with softmax activation to output class probabilities. Let C represent the number of emotion classes (here, $C = 6$). The output vector is given by:

$$\mathbf{y_i} = \text{Softmax}(\mathbf{W} \cdot \mathbf{z}),$$

where $\mathbf{W} \in \mathbb{R}^{C\times18}$ is the weight matrix, and $C = 6$ corresponds to the six emotion classes. The resulting output $\mathbf{y_i}$ is a real-valued vector in $\mathbb{R}^C$. Each component $\mathbf{y_i}$ represents the predicted probability of the input belonging to class i.

D. Transfer Learning

The proposed model leverages TL to achieve high accuracy in emotion recognition while maintaining a lightweight architecture with $\approx$ 28K parameters. The TL process involves pretraining the proposed model on HAR data before retraining it on HER data. The HAR data was segmented into fixed-length windows of 256 samples. The model was initialized with random weights and trained for 40 epochs using categorical cross-entropy loss, the Adam optimizer, and accuracy as the evaluation metric. This pretraining phase enabled the model to learn general temporal patterns and motion-related features. For the HER task, the model's final layer, originally designed for activity classification, was replaced with a

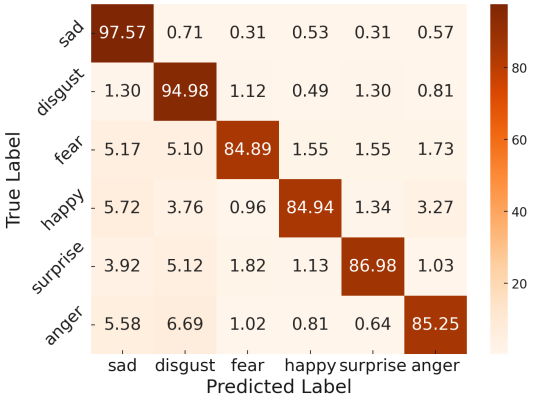


Fig. 4: Confusion Matrix for 1D Input ($smartphone \xrightarrow[\text{retrained with}]{} smartphone$): The model was pretrained using 1D HAR data collected from a smartphone and then retrained with 1D HER data, also collected from a smartphone.

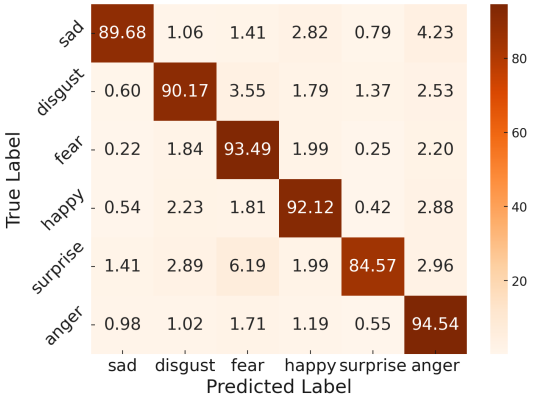


Fig. 5: Confusion Matrix for 1D Input ($smartwatch \xrightarrow[\text{retrained with}]{} smartphone$): The model was pretrained using 1D HAR data collected from a smartwatch and then retrained with 1D HER data collected from a smartphone.

softmax layer containing six nodes for emotion classification. The HER data was similarly segmented into windows of 256 samples. The model was then retrained with all layers for 90 epochs. This ensured the adaptation of the features learned during pretraining to the emotion recognition task. Algorithm 1 presents the exact details of the entire process.

IV. Results

To evaluate the performance of the proposed model, we conduct training and validation both with and without pretrained knowledge. The results of these experiments are discussed below.

A. Benchmarking with Pretrained Knowledge

Here, we present the benchmarking results using pretrained knowledge for different input sizes, including 1D, 3D, and 6D.

1) A Minimalistic Approach (1D Input): The minimalistic approach results (confusion matrices), i.e., 1D input computed from 3D accelerations, are presented in Figures 4 and 5. The best performance is observed when the model is pretrained with smartphone HAR data and subsequently retrained with smartphone HER data ($smartphone \xrightarrow[\text{retrained with}]{} smartphone$). Figure 6 displays the precision, recall, and F1-scores for all different use cases.

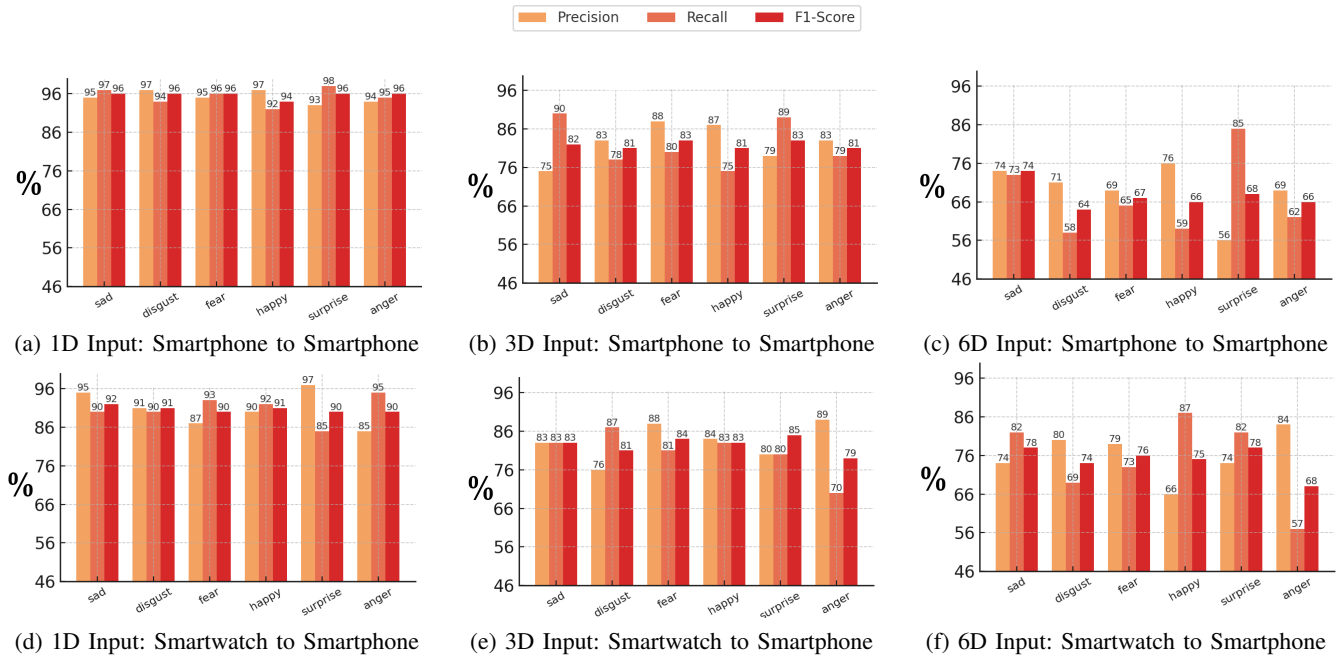
For the class "sad", the phone model achieved 95% precision, 97% recall, and 96% F1-score, outperforming the watch model, which achieved 95% precision, 90% recall, and 92% F1-score. A similar trend was observed for "disgust", where the phone model reached 97% precision and 94% recall, 96% F1-score, compared to the watch's 91% precision and 90% recall, 91% F1-score. In the case of "fear", the phone again led with 95% precision, 96% recall, 96% F1-score, while the watch scored 87% precision, 93% recall, 90% F1-score.

For the class "happy", the phone model achieved 97% precision, 92% recall, 94% F1-score, surpassing the watch's 90% precision, 92% recall, 91% F1-score. However, "surprise" presented an interesting contrast: while the phone scored 93% precision, 98% recall, and 96% F1-score, the watch achieved higher precision 97% but lower recall 85%, resulting in a 90% F1 score. Finally, both models handled "anger" well, though the phone model, with 94% precision, 95% recall, and 96% F1-score, slightly outperformed the watch, which achieved 85% precision, 95% recall, and 90% F1-score.

Overall, these differences resulted in 95% accuracy for the phone-based 1D input and 91% accuracy for the watch-based 1D input. This indicated that the phone-based 1D input for pretraining was more effective. In contrast, pretraining with watch-based 1D input was less effective. We hypothesize that since the emotions data is also from a phone, which stays attached to the same location, the pre-learned features from the phone were more suitable for TL while the watch was worn on the dominant hand, which may have introduced additional arm-centric variability, thereby impacted its performance.

2) Leveraging Accelerometer Data (3D Input): The confusion matrix in Figure 7 corresponds to the model pretrained using 3D HAR data from a smartwatch and subsequently retrained with 3D HER data from a smartphone ($smartwatch \xrightarrow[\text{retrained with}]{} smartphone$). Similarly, the confusion matrix in Figure 8 represented the model pretrained with 3D HAR data from a smartphone and then retrained with 3D HER data from a smartphone ($smartphone \xrightarrow[\text{retrained with}]{} smartphone$). A performance comparison between the two models revealed that the former outperformed the latter across most metrics. Specifically, it achieved higher F1-scores for "sad" (83% vs. 82%), "fear" (84% vs. 83%), "happy" (83% vs. 81%), and "surprise" (85% vs. 83%). Overall, the former model attained a higher macro-average F1-score of 83%, compared to 82% for the latter, indicating slightly better generalization across all emotions. Figure 6 provides a detailed breakdown of precision, recall, and F1-scores for different use cases.

3) Integrating Accelerations and Angular Velocities (6D Input): A significant degradation in the performance of the model was observed when a 6D input having integrated accelerations and angular velocities was used for pretraining using HAR and retraining using HER datasets. Figures 9, 10



(a) 1D Input: Smartphone to Smartphone (b) 3D Input: Smartphone to Smartphone (c) 6D Input: Smartphone to Smartphone

(d) 1D Input: Smartwatch to Smartphone (e) 3D Input: Smartwatch to Smartphone (f) 6D Input: Smartwatch to Smartphone

Fig. 6: Comparison of Precision, Recall, and F1-score for different input dimensions and device transfer scenarios.

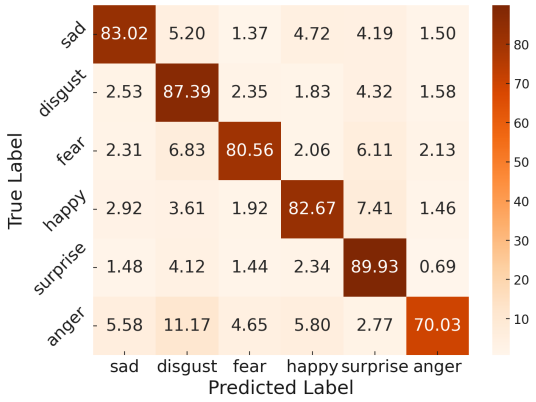


Fig. 7: Confusion Matrix for 3D Accelerations ($smartwatch \xrightarrow[\text{retrained with}]{} smartphone$): The model was pretrained using 3D HAR data collected from a smartwatch and then retrained with 3D HER data collected from a smartphone.

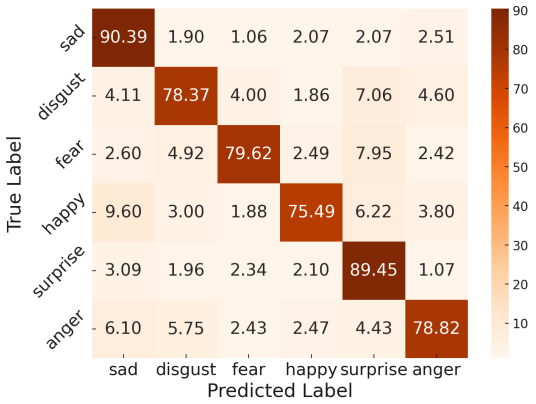


Fig. 8: Confusion Matrix for 3D Accelerations ($smartphone \xrightarrow[\text{retrained with}]{} smartphone$): The model was pretrained using 3D HAR data collected from a smartphone and then retrained with 3D HER data collected from a smartphone.

showed confusion matrices for phone and watch, respectively. When pretrained with smartwatch HAR data, the average classification accuracy remained at 75.04%, which dropped to 67.29% when smartphone HAR data was used in pretraining. The F1-scores for the former case remained at 78% for "sad", 74% for "disgust", 76% for "fear", 75% for "happy", 78% for "surprise", and 68% for "anger". For the latter case, the F1-scores remained at 74%, 64%, 67%, 66%, 68%, and 66%, respectively. Details can be found in Figure 6.

B. Evaluating the Model Without Pretrained Knowledge

We also computed the results of the model without pretrained knowledge for different input sizes, including 1D, 3D, and 6D. The best performance was achieved with a 1D input size. The details are provided below.

1) Minimal Input (1D): For the 1D input, without TL, the classifier achieved 85% accuracy, with macro- and weighted-average F1-scores of 85% (Figure 12). "Sad" is well-distinguished, achieving 91% precision, recall, and F1-score. "Disgust" has high recall (90%) but lower precision 79%, leading to some misclassifications. "Fear" achieves 86% precision, 83% recall, and an 85% F1-score. Both "happy" and "anger" reach 85% F1-scores, with 83% precision and 88% recall. "Surprise" stands out with 91% precision but lower recall 73%, resulting in many true cases being missed.

The confusion matrix Figure 13 confirms these trends: "Sad" is correctly classified 91.27% of the time, with slight confusion toward "happy" 3.88%. "Disgust" is correctly identified 89.82% of the time but is occasionally misclassified as "anger" 3.02% or "happy" 2.53%. "Fear" achieves 83.34%

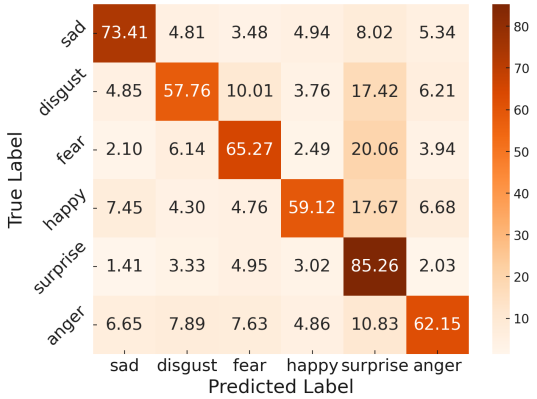


Fig. 9: Confusion Matrix for 6D Accelerations and Angular Velocities ($smartphone \xrightarrow[\text{retrained with}]{} smartphone$): The model was pretrained using 6D HAR data collected from a smartphone and then retrained with 3D HER data collected from a smartphone.

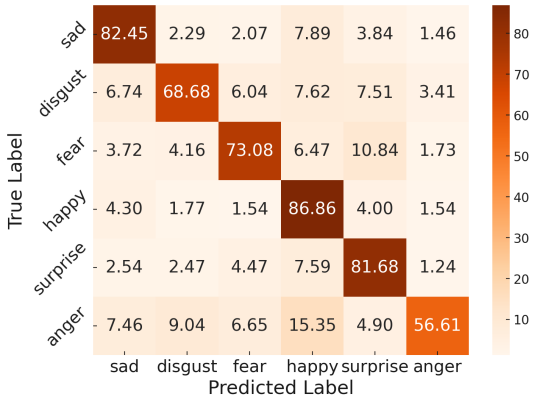
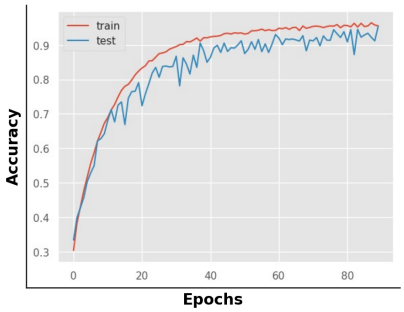


Fig. 10: Confusion Matrix for 6D Accelerations and Angular Velocities ($smartwatch \xrightarrow[\text{retrained with}]{} smartphone$): The model was pretrained using 6D HAR data collected from a smartwatch and then retrained with 3D HER data collected from a smartphone.

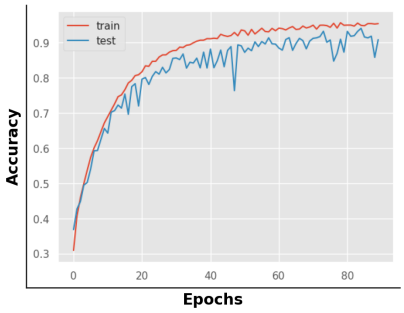
accuracy, overlapping with "anger" 5.53% and "surprise" 4.19%. "Happy" is correctly classified 87.74% of the time, with some misclassifications as "anger" 3.38%. "Surprise" has 72.71% accuracy but is often misclassified as "disgust" 14.50% or "fear" 5.95%. Finally, "anger" is correctly classified 87.94% of the time but overlaps with "happy" 5.88% and "fear" 1.53%.

2) Accelerometer Data (3D Input): Using only 3D accelerometer data, the model achieved an overall accuracy of 86.77%, along with a macro-average F1-score of 87%. Its weighted-average precision, recall, and F1-score were also around 87%, as shown in Figure 12, indicating that even without gyroscope inputs, acceleration signals alone captured key gait dynamics for emotion classification.

In terms of individual classes, "sad" performed with 88% precision, 89% recall, and an 89% F1-score, showing strong detection and few false positives. "Disgust" stood out with a high precision of 92% but a recall of only 81%, suggesting that while most of its positive predictions were correct, the model occasionally missed true disgust samples. "Fear" demonstrated 86% precision and 90% recall, whereas "happy" showed 87%



(a) Phone: Magnitude of 3D Acceleration



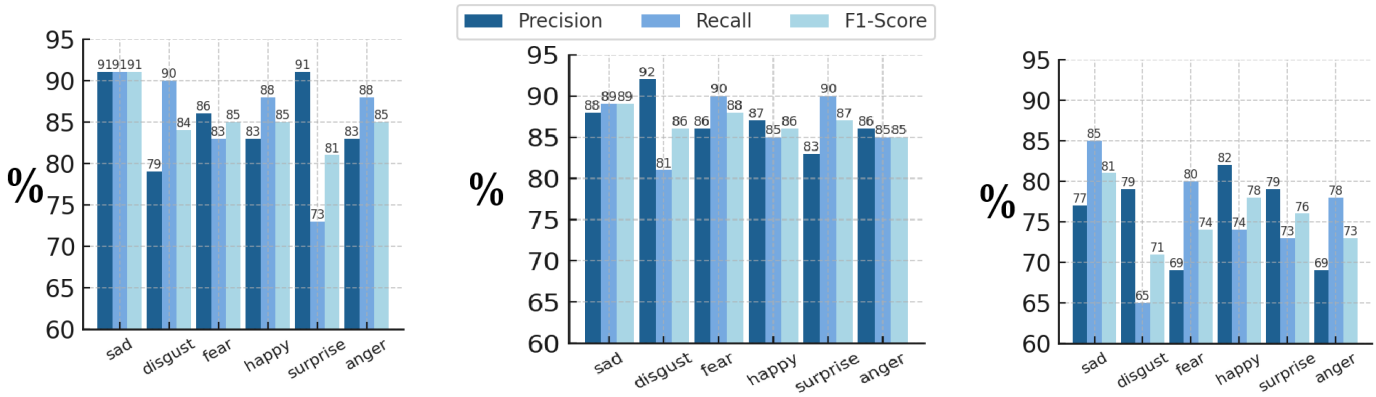
(b) Watch: Magnitude of 3D Acceleration

Fig. 11: The accuracy plots, computed using pretrained knowledge, illustrate (a) the smartphone's 3D accelerations and (b) the smartwatch's 3D accelerations. The model converges well, effectively ruling out overfitting.

precision and 85% recall; both classes were reliably detected but were prone to minor misclassifications. "Surprise" had 83% precision and a recall of 90%, meaning it was captured in most cases but occasionally confused with other classes. Meanwhile, "anger" attained 86% precision and 85% recall, indicating balanced performance with some overlap.

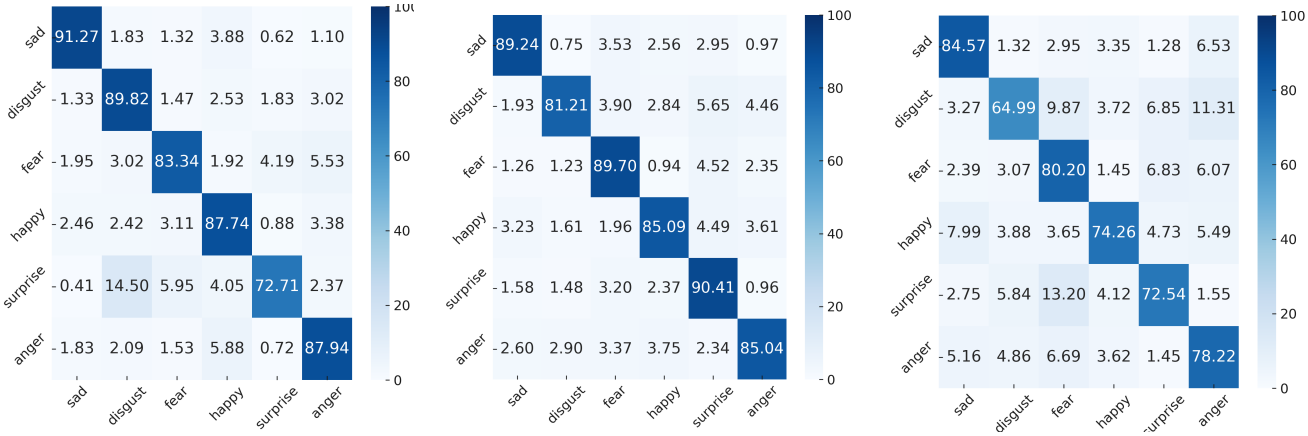
The confusion matrix (Figure 13) reaffirmed these findings. "Sad" was correctly recognized 89.24% of the time, with slight misclassification toward "fear" and "happy". "Disgust" was correct 81.21% of the time, but about 5.65% of its data were labeled as "surprise" and 4.46% as "anger", reflecting lower recall. "Fear" showed 89.70% accuracy, often overlapping with "surprise" and "anger", while "happy" achieved 85.09%, mixing with "sad" and "anger" in some instances. "Surprise" had the highest correctness at 90.41% but occasionally confused other classes. "Anger" was correct 85.04% of the time, with moderate spillovers into "fear" and "happy".

3) Combining Accelerations and Angular Velocities (6D Input): Using 6D inertial data, accelerometer and gyroscope, the model achieved an overall accuracy of 75.38%, with a macro-average F1-score of 75% and weighted-average precision and recall also near 75%. Class-wise performance, as shown in Figure 12, revealed that "sad" was recognized with 77% precision, 85% recall, and an 81% F1-score, indicating that the model was particularly adept at identifying sad instances, albeit occasionally over-predicting them. "Disgust" had 79% precision but only 65% recall, suggesting that while many of its positive predictions were correct, it often failed to detect all disgust examples. "Fear" exhibited 69% precision



(a) Input size: 1D (b) Input size: 3D (c) Input size: 6D

Fig. 12: Evaluation of the proposed model without pretrained knowledge. Each bar represents a metric (Precision, Recall, or F1-Score) for a given category. The best results are seen for the 1D input.



(a) 1D Without TL (b) 3D Without TL (c) 6D Without TL

Fig. 13: Confusion matrices computed without pretrained knowledge for different input sizes (1D, 3D, and 6D). The best results are observable for the 1D input size.

and 80% recall, meaning it was accurately captured most of the time but at the cost of some false positives. "Happy" demonstrated one of the highest precisions at 82% and a recall of 74%, revealing that truly happy samples were sometimes misclassified. "Surprise" attained 79% precision, 73% recall, and a 76% F1-score, reflecting moderate confusion with fear and happy. Meanwhile, "anger" achieved 69% precision and 78% recall, indicating a decent capture rate but moderate confusion with other emotions.

The confusion matrix Figure 13 revealed that "sad" was correctly recognized 84.57% of the time, with some leakage into "anger" 6.53% and "fear" 2.95%. "Disgust" saw a true positive rate of 64.99% but was confused with "fear" 9.87% and "anger" 11.31%. "Fear" was correctly classified 80.20% of the time yet overlapped occasionally with "surprise" and "anger", while "happy" was identified in 74.26% of the corresponding samples but could be misread as "sad" around 8% or "anger" 5.49%. "Surprise" was recognized about 72.54% of the time, with its largest confusion arising from "fear" 13.20%. Finally, "anger" showed a correct classification rate of 78.22% but was sometimes mislabeled as "fear" or "sad".

These misclassifications suggested that certain emotional states such as fear and surprise exhibited overlapping gait characteristics, and that disgust may have shared more subtle similarities with anger or fear, thus making the classification without pretrained knowledge challenging.

C. On-Device Inference Evaluation

To evaluate the on-device feasibility of our proposed model, we benchmarked the inference performance of Jazbat-Net against the previous SOTA model, using TensorFlow Lite (TFLite) in FP32 precision. We intentionally avoided quantization to ensure numerical consistency with the GPU-powered VM inference results.

The experiments were conducted on a Xiaomi 13T Global Version smartphone (SoC: MediaTek Dimensity 8200-Ultra, Android 15, Kernel 5.10.226) using the Benchmark Model Plus Flex tool provided by TFLite, executed via the ADB interface. All evaluations were performed on the CPU backend with XNNPACK acceleration enabled. The input tensor was of size $1 \times 256 \times 1$, representing the magnitude of acceleration (1D input). To ensure consistency and reproducibility, we performed 5 warm-up runs followed by 100 timed runs.

The results demonstrate that Jazbat-Net achieves a median latency of 90.96 ms and a mean latency of 90.37 ± 7.94 ms, with latencies ranging from 68.62 ms to 110.54 ms. The size of the TFLite FP32 model is only 0.158 MB. In comparison, the previous SOTA model yields a median latency of 111.86 ms and a mean latency of 113.16 ± 9.72 ms, with latencies between 92.19 ms and 151.20 ms, and a TFLite FP32 model size of 3.11 MB.

TABLE I: On-device inference performance comparison between Jazbat-Net and the Previous SOTA model on a Xiaomi 13T smartphone (CPU backend, FP32, XNNPACK)

Model	FP32 File Size (MB)	Mean $\pm$ SD Latency (ms)	Min / Max Latency (ms)
Jazbat-Net	**0.158**	90.37 ± 7.94	**68.62 / 110.54**
SOTA [9]	3.11	113.16 ± 9.72	92.19 / 151.20

Overall, Jazbat-Net delivers a $\approx 19\%$ reduction in median latency and a $\approx 20\times$ reduction in model size, while maintaining full FP32 numerical accuracy. A detailed comparison is provided in Table I.

D. Comparison with Previous Studies

Table II compares our emotion classification model with previous studies. Most prior works employed traditional machine learning models such as SVM, RF, DT, and MLP for emotion classification, often focusing on a limited set of emotions. For instance, Piskioulis et al. [34] classified enjoyment and frustration with approximately 89% accuracy, while Hashmi et al. [31] classified six emotions with 86.45% accuracy. Some studies, such as Imran et al. [9], achieved a high accuracy 95.23%; however, their model did not employ TL and their trainable parameters remained above 700K. Our study presents a lightweight model with only $\approx$ 28K parameters, achieving 95.36% accuracy on six emotions. Unlike previous works, we employ pretraining on a HAR dataset, followed by retraining on the emotion recognition dataset. The impact of pretraining is evident; models trained without pretraining (e.g., Hamza et al. [10]) achieved 78.20% accuracy, whereas models pretrained on acceleration magnitude ($\widehat{mag_a^p}$) reached 95.36%, demonstrating the effectiveness of pretrained knowledge and TL. These results highlight that an optimized, compact model, when pretrained effectively, can achieve state-of-the-art accuracy while remaining computationally efficient. This makes our approach highly suitable for real-time and resource-constrained applications.

E. Complexity Analysis

We evaluate and report the efficiency of Jazbat-Net from three perspectives: (i) theoretical time complexity ($\mathcal{O}$) derived from the model's architecture, (ii) Multiply–Accumulate Computations (MACs) to quantify the actual computational workload, and (iii) space complexity in terms of parameter count and memory usage. Furthermore, we compare Jazbat-Net against the SOTA model [9].

1) Time Complexity : For a sequence of length *T* (which represents the number of time steps in an input segment after segmentation; T=256 in all experiments), input dimension *d*, and hidden dimension *h*, the time complexity of the BiGRU component in the proposed model is:

$$\mathcal{O}\big(T \cdot (dh + h^2)\big) \tag{2}$$

For the CNN component, let the kernel size be *k*, the number of input channels be C_{in}, the number of output channels be C_{out}, the feature map length be *T*, and the number of convolutional layers be *L*. The time complexity can be expressed as:

$$\mathcal{O}\big(T \cdot C_{\text{in}} \cdot C_{\text{out}} \cdot k \cdot L\big) \tag{3}$$

For the fully connected layer of size $h \times n_{\text{class}}$, the time complexity is:

$$\mathcal{O}\big(h \cdot n_{\text{class}}\big) \tag{4}$$

Therefore, the overall time complexity of the proposed *Jazbat-Net* model is:

$$\mathcal{O}\Big(T \cdot (dh + h^2) + T \cdot C_{\text{in}} \cdot C_{\text{out}} \cdot k \cdot L + h \cdot n_{\text{class}}\Big) \tag{5}$$

2) MACs-based Complexity : MACs quantify the number of multiplications and additions required to generate the model output. Building on the time complexity analysis, we computed the total MACs for different input feature dimensions and reported the results for both Jazbat-Net and the SOTA model in Table III.

Across varying input configurations ($F \in \{1, 3, 6\}$), Jazbat-Net consistently requires approximately 22–23$\times$ fewer MACs than the SOTA model. This corresponds to an average reduction of about 95.5% in computational cost per forward pass.

3) Space Complexity : The space complexity of Jazbat-Net primarily arises from three components: (i) the BiGRU layers, (ii) the multi-scale CNN modules, and (iii) the fully connected layer.

$$\mathcal{O}\Big(T \cdot (dh + h^2) + T \cdot C_{\text{in}} \cdot C_{\text{out}} \cdot k \cdot L + h \cdot n_{\text{class}}\Big) \tag{6}$$

For $T = 256$, $h = 32$, $L = 3$, and $n_{\text{class}} = 6$, the detailed calculation of trainable parameters and activation memory for Jazbat-Net is summarized in Table IV. The BiGRU contributes approximately 6,720 trainable parameters with an activation memory of $\mathcal{O}(2Th) = 16{,}384$. Batch Normalization (BN) has 256 parameters in total, out of which 128 are trainable, with negligible activation memory. The three multi-scale CNN modules contribute around 20,952 parameters with an activation memory of $\mathcal{O}(T \cdot C_{\text{out}} \cdot k) = 10{,}240$. The size-reduction 1×1 convolution adds 390 parameters with ≈ 256 activations. The fully connected (FC) layer contributes 42 parameters with an activation memory of only 6 (equal to the number of classes).

TABLE II: Performance Comparison with Previous Studies

Sr. #	Ref.	Emotions Classified	Model Type	Accuracy (%)
1	Piskioulis et al. ([34])	Enjoyment and Frustration	LR, DT, LR, SVM, MLP	Enjoyment (87.90), Frustration (89.45)
2	Reyana et al. ([35])	Neutral, Angry, happy and sad	ML	Neutral (100), Angry (90), happy (80) and sad (70)
3	Hashmi et al. ([31])	happy, fear, sad, disgust, anger and surprise	SVM, RF	86.45
4	Quiroz et al. ([36])	happy, sad	LR, RF	75
5	Cui et al. ([37])	happy, anger and Neutral	SVM, DT, MLP, RT, RF	80
6	Zhang et al. ([38])	happy, anger and Neutral	SVM, DT, RT, RF	81.2
7	Hamza et al. [10] (No Pretraining)	happy, fear, sad, disgust, anger & surprise	$a_x^p, a_y^p, a_z^p, \omega_x^w, \omega_y^p, \omega_z^p$	78.198
8	Imran et al. [9] (No Pretraining)	happy, fear, sad, disgust, anger & surprise	$(\widehat{mag_a^p})$	95.23
9	Pre-Training:$(a_x^p, a_y^p, a_z^p, \omega_x^p, \omega_y^p, \omega_z^p)$	happy, fear, sad, disgust, anger & surprise	Post-Training: $a_x^p, a_y^p, a_z^p, \omega_x^w, \omega_y^p, \omega_z^p$	67.29
10	Pre-Training:$(\widehat{a_x^p, a_y^p}, a_z^p)$	happy, fear, sad, disgust, anger & surprise	Post-Training: $\widehat{a_x^p, a_y^p}, a_z^p$	81.96
11	Pre-Training: $(\widehat{mag_a^p})$	happy, fear, sad, disgust, anger & surprise	Post-Training: $\widehat{mag_a^p}$	95.36
12	Pre-Training:$(a_x^w, a_y^w, a_z^w, \omega_x^w, \omega_y^w, \omega_z^w)$	happy, fear, sad, disgust, anger & surprise	Post-Training: $a_x^p, a_y^p, a_z^p, \omega_x^p, \omega_y^p, \omega_z^p$	75.04
13	Pre-Training:(a_x^w, a_y^w, a_z^w)	happy, fear, sad, disgust, anger & surprise	Post-Training: a_x^p, a_y^p, a_z^p	82.66
14	Pre-Training:$(\widehat{mag_a^w})$	happy, fear, sad, disgust, anger & surprise	Post-Training: $(\widehat{mag_a^p})$	90.60

TABLE III: Computational cost comparison between Jazbat-Net and previous SOTA [9].

Model	Input (F)	MACs
Jazbat-Net	1D	6.96M
	3D	7.05M
	6D	7.20M
SOTA [9]	1D	157.88M
	3D	158.14M
	6D	158.53M

TABLE IV: Space complexity of Jazbat-Net (trainable parameters and activation memory).

	Parameters		Activation Memory
	Trainable	Total	($\approx$)
BiGRU ($L = 3$, $h = 32$)	6,720	6,720	16,384
BatchNorm (γ, β)	128	256	Negligible
3 Multi-scale CNN Modules	20,952	20,952	10,240
Size Reduction 1×1 Conv	390	390	≈ 256
Fully Connected Layer	42	42	≈ 6
Total	**28,232**	**28,360**	$\approx$ 26,886

TABLE V: Comparison of model size between Jazbat-Net and previous SOTA [9] across different input configurations.

Model	Input (F)	Parameters	Size (FP32)
Jazbat-Net	1D	$\approx$ 28.36K	$\approx$ 110 KB
	3D	$\approx$ 28.55K	$\approx$ 113 KB
	6D	$\approx$ 29.13K	$\approx$ 116 KB
SOTA [9]	1D	$\approx$ 754K	$\approx$ 3.0 MB
	3D	$\approx$ 756K	$\approx$ 3.0 MB
	6D	$\approx$ 757K	$\approx$ 3.1 MB

In total, Jazbat-Net has approximately 28.23K trainable parameters and a total parameter count of $\approx$ 28.36K, including non-trainable BN statistics. The activation memory requirement is around 26.9K, dominated by the BiGRU and CNN modules. Compared to the previous SOTA [9], which uses deeper BiGRUs and wider CNNs, Jazbat-Net achieves a significantly lower parameter count and reduced space complexity, as summarized in Table V.

During inference, Jazbat-Net requires $\approx 0.25-0.35\,\mathrm{MB}$ of activation memory due to its efficient design. For an input window of $T = 256$ samples and a batch size of 1, the largest intermediate tensors arise in the three residual inception-like modules, each handling feature maps up to 256×64. Across all branches and concatenations, this corresponds to approximately $54{,}272$ floats, or $\approx 0.21\,\mathrm{MB}$ in FP32 precision. Adding the BiGRU outputs, batch normalization buffers, and small activations from the 1×1 projections and the final dense layer, the peak activation memory remains within $\approx 0.25-0.35\,\mathrm{MB}$.

In contrast, the previous SOTA model consumes $\approx 3-4\,\mathrm{MB}$ due to its deeper BiGRUs and wider convolutional layers, leading to significantly larger intermediate tensors. Overall, Jazbat-Net is $\approx 26\times$ smaller in model size and uses $\approx 10-12\times$ less activation memory, making it highly suitable for real-time IoT and wearable applications where a low memory footprint and fast inference are critical.

TABLE VI: Performance comparison for different all cases: no pre-training, pre-training with watch data, and pre-training with phone data.

Input Size	Without pre-training (%)	Pretrained using (%)	
		Smartwatch	Smartphone
6D (a_x, a_y, a_z $\omega_x, \omega_y, \omega_z$)	75.38	75.04	67.29
3D (a_x, a_y, a_z)	86.77	82.66	81.96
1D ($\widehat{mag_a}$)	**85.10**	**90.62**	**95.36**

V. Conclusion

This study highlights the potential of TL in bridging HAR and human emotion recognition using wearable inertial sensor data. We introduce Jazbat-Net, a lightweight and efficient deep learning model built on the principles of TL, effectively tackling the challenge of limited labeled datasets for emotion recognition. The proposed model is first pretrained on a large-scale, publicly available multi-activity HAR dataset and then retrained on an HER dataset. The model achieved state-of-the-art accuracy with significantly reduced complexity, making it highly suitable for resource-constrained IoT applications.

The performance of the proposed model has been evaluated both with and without TL, using different input sizes (1D, 3D, and 6D), as shown in Table VI. The best results were achieved with pretrained knowledge and a 1D input size, where an average classification accuracy of over 95% was obtained. Compared to previous studies summarized in Table II, our work is the most efficient and includes the largest number of emotional states. Furthermore, to rule out overfitting, accuracy vs. epoch curves were plotted, as shown in Figure 11.

The findings indicate that TL from HAR to HER is effective, particularly when the source and target data are aligned in terms of sensor placement. Simpler inputs, such as 1D

acceleration magnitude, proved to be more robust and transferable than higher-dimensional data. The performance drop observed with 6D inputs suggests that added complexity can introduce noise rather than improve results. Similarly, phone-based pretraining outperformed watch-based pretraining due to more consistent placement and reduced motion variability. These results underscore the value of low-complexity, domain-aligned features in emotion recognition. Future work should address real-world variability and explore multimodal sensing, unsupervised learning, and model interpretability.

A. Limitations and Future Directions

Although the results are promising, several limitations remain that warrant further investigation. These include constraints related to dataset size, diversity, and the challenges of real-world deployment. Future research directions could address these limitations by incorporating multi-modal sensor data, conducting cross-cultural and longitudinal studies, and leveraging self-supervised learning to reduce reliance on labeled data. Moreover, advancing explainability in emotion recognition models and optimizing them for deployment in real-world IoT environments may enable broader adoption and greater practical impact.

REFERENCES

[1] A. A. Zakaria, T. Amr, and A. A. Ragheb, "Iot in smart urban planning: A comprehensive review of applications, developments, and engineering perspectives," *IEEE Access*, vol. 13, pp. 135 316–135 335, 2025.

[2] I. Ahmad, Z. Asghar, T. Kumar, G. Li, A. Manzoor, K. Mikhaylov, S. A. Shah, M. Höyhtyä, J. Reponen, J. Huusko, and E. Harjula, "Emerging technologies for next generation remote health care and assisted living," *IEEE Access*, vol. 10, pp. 56 094–56 132, 2022.

[3] O. Taiwo and A. E. Ezugwu, "Internet of things-based intelligent smart home control system," *Security and Communication Networks*, vol. 2021, no. 1, p. 9928254, 2021.

[4] S. K. Challa, A. Kumar, and V. B. Semwal, "A multibranch cnn-bilstm model for human activity recognition using wearable sensor data," *The Visual Computer*, vol. 38, no. 12, pp. 4095–4109, 2022.

[5] X. Yu, J. Jang, and S. Xiong, "A large-scale open motion dataset (kfall) and benchmark algorithms for detecting pre-impact fall of the elderly using wearable inertial sensors," *Frontiers in Aging Neuroscience*, vol. 13, p. 692865, 2021.

[6] G. Grouvel, L. Carcreff, F. Moissenet, and S. Armand, "A dataset of asymptomatic human gait and movements obtained from markers, imus, insoles and force plates," *Scientific Data*, vol. 10, no. 1, p. 180, 2023.

[7] Y. Luo, S. M. Coppola, P. C. Dixon, S. Li, J. T. Dennerlein, and B. Hu, "A database of human gait performance on irregular and uneven surfaces collected by wearable sensors," *Scientific data*, vol. 7, no. 1, p. 219, 2020.

[8] C. Moreau, T. Rouaud, D. Grabli, I. Benatru, P. Remy, A.-R. Marques, S. Drapier, L.-L. Mariani, E. Roze, D. Devos *et al.*, "Overview on wearable sensors for the management of parkinson's disease," *npj Parkinson's Disease*, vol. 9, no. 1, p. 153, 2023.

[9] H. A. Imran, Q. Riaz, M. Zeeshan, M. Hussain, and R. Arshad, *Applied Sciences*, vol. 13, no. 8, p. 4728, 2023.

[10] K. Hamza, Q. Riaz, H. A. Imran, M. Hussain, and B. Krüger, "Generisch-net: A generic deep model for analyzing human motion with wearable sensors in the internet of health things," *Sensors*, vol. 24, no. 19, p. 6167, 2024.

[11] H. A. Imran, K. Hamza, and Z. Mehmood, "Harresnext: An efficient resnext inspired network for human activity recognition with inertial sensors," in *2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, 2022, pp. 1–4.

[12] I. B. Mauss and M. D. Robinson, "Measures of emotion: A review," *Cognition and emotion*, vol. 23, no. 2, pp. 209–237, 2009.

[13] B. M. S. Inguscio, G. Cartocci, S. Palmieri, S. Menicocci, A. Vozzi, A. Giorgi, S. Ferrara, P. Canettieri, and F. Babiloni, "Poetry in pandemic: A multimodal neuroaesthetic study on the emotional reaction to the divina commedia poem," *Applied Sciences*, vol. 13, no. 6, p. 3720, 2023.

[14] L. Chen, K. Wang, M. Li, M. Wu, W. Pedrycz, and K. Hirota, "K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition in human–robot interaction," *IEEE Transactions on Industrial Electronics*, vol. 70, no. 1, pp. 1016–1024, 2023.

[15] Y. Zhao, M. Guo, X. Sun, X. Chen, and F. Zhao, "Attention-based sensor fusion for emotion recognition from human motion by combining convolutional neural network and weighted kernel support vector machine and using inertial measurement unit signals," *IET signal processing*, vol. 17, no. 4, p. e12201, 2023.

[16] O. Napoli, D. Duarte, P. Alves, D. H. P. Soto, H. E. de Oliveira, A. Rocha, L. Boccato, and E. Borin, "A benchmark for domain adaptation and generalization in smartphone-based human activity recognition," *Scientific Data*, vol. 11, no. 1, p. 1192, 2024.

[17] M. Pesenti, G. Invernizzi, J. Mazzella, M. Bocciolone, A. Pedrocchi, and M. Gandolla, "Imu-based human activity recognition and payload classification for low-back exoskeletons," *Scientific Reports*, vol. 13, no. 1, p. 1184, 2023.

[18] R. Raj and A. Kos, "An improved human activity recognition technique based on convolutional neural network," *Scientific Reports*, vol. 13, no. 1, p. 22581, 2023.

[19] M. Khan and Y. Hossni, "A comparative analysis of lstm models aided with attention and squeeze and excitation blocks for activity recognition," *Scientific Reports*, vol. 15, no. 1, p. 3858, 2025.

[20] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[21] T. Pandit, H. Nahane, D. Lade, and V. Rao, "Abnormal gait detection by classifying inertial sensor data using transfer learning," in *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2019, pp. 1444–1447.

[22] L. Pei, S. Xia, L. Chu, F. Xiao, Q. Wu, W. Yu, and R. Qiu, "Mars: Mixed virtual and real wearable sensors for human activity recognition with multidomain deep learning model," *IEEE Internet of Things Journal*, vol. 8, no. 11, pp. 9383–9396, 2021.

[23] P.-H. Kuo, Y.-C. Shen, P.-H. Feng, Y.-J. Chiu, and H.-T. Yau, "Transfer learning-based gesture and pose recognition system for human robot interaction: An internet of things application," *IEEE Internet of Things Journal*, 2024.

[24] Z. Fu, X. He, E. Wang, J. Huo, J. Huang, and D. Wu, "Personalized human activity recognition based on integrated wearable sensor and transfer learning," *Sensors*, vol. 21, no. 3, p. 885, 2021.

[25] J. Link, T. Perst, M. Stoeve, and B. M. Eskofier, "Wearable sensors for activity recognition in ultimate frisbee using convolutional neural networks and transfer learning," *Sensors*, vol. 22, no. 7, p. 2560, 2022.

[26] Y. Celik, M. F. Aslan, K. Sabanci, S. Stuart, W. L. Woo, and A. Godfrey, "Improving inertial sensor-based activity recognition in neurological populations," *Sensors*, vol. 22, no. 24, p. 9891, 2022.

[27] Y. Zhang, Y. Shao, R. Luo, L. Xiong, and J. Zhang, "Multiple human activities classification based on dynamic on-body propagation characteristics using transfer learning," *IEEE Internet of Things Journal*, 2023.

[28] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.

[29] B. Zhang, R. Zheng, and J. Liu, "A multi-source unsupervised domain adaptation method for wearable sensor based human activity recognition," in *Proceedings of the 20th International Conference on Information Processing in Sensor Networks (co-located with CPS-IoT Week 2021)*, 2021, pp. 410–411.

[30] G. M. Weiss, K. Yoneda, and T. Hayajneh, "Smartphone and smartwatch-based biometrics using activities of daily living," *IEEE Access*, vol. 7, pp. 133 190–133 202, 2019.

[31] M. A. Hashmi, Q. Riaz, M. Zeeshan, M. Shahzad, and M. M. Fraz, "Motion reveal emotions: Identifying emotions from human walk using chest mounted smartphone," *IEEE Sensors Journal*, vol. 20, no. 22, pp. 13 511–13 522, 2020.

[32] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[33] H. A. Imran, Q. Riaz, M. Hussain, H. Tahir, and R. Arshad, "Smart-wearable sensors and cnn-bigru model: A powerful combination for human activity recognition," *IEEE Sensors Journal*, vol. 24, no. 2, pp. 1963–1974, 2023.

[34] O. Piskioulis, K. Tzafilkou, and A. Economides, "Emotion detection through smartphone's accelerometer and gyroscope sensors," in *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, 2021, pp. 130–137.
[35] A. Reyana, P. Vijayalakshmi, and S. Kautish, "Multisensor fusion approach: a case study on human physiological factor-based emotion recognition and classification," *International Journal of Computer Applications in Technology*, vol. 66, no. 2, pp. 107–114, 2021.
[36] J. C. Quiroz, M. H. Yong, and E. Geangu, "Emotion-recognition using smart watch accelerometer data: Preliminary findings," in *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, 2017, pp. 805–812.
[37] L. Cui, S. Li, and T. Zhu, "Emotion detection from natural walking," in *International Conference on Human Centered Computing*. Springer, 2016, pp. 23–33.
[38] Z. Zhang, Y. Song, L. Cui, X. Liu, and T. Zhu, "Emotion recognition based on customized smart bracelet with built-in accelerometer," *PeerJ*, vol. 4, p. e2258, 2016.