HUMAPS-4D : A Multimodal Dataset for HUman Motion Analysis with Physiological and Semantic informations

Supplementary Material

6. 3D pose estimation from foot pressure

Given a temporal sequence of insole measurements, our objective is to infer the full set of 24 3D joints positions recorded by a Qualisys Track Manager system. At each time step t, the plantar pressure vector is denoted

$$\mathbf{P}_t \in \mathbb{R}^{2N_s},$$

where each foot contains N_s pressure sensors, while the auxiliary insole vector

$$\mathbf{A}_t \in \mathbb{R}^{18}$$

includes tri-axial accelerations, foot orientation (pitch, yaw, roll), center-of-pressure coordinates (x, y) for each foot, and total ground–reaction forces (left and right). The model (see Figure 6) learns a mapping

$$f : (\mathbf{P}_{1:T},\ \mathbf{A}_{1:T}) \longrightarrow \mathbf{J}_{1:T}, \tag{7}$$

where $\mathbf{J}_t \in \mathbb{R}^{72}$ contains the 3D positions of the 24 joints at time t. Figure 5 illustrates the layout of the indidivual pressure sensors in the Moticon OpenGo instrumented insoles.

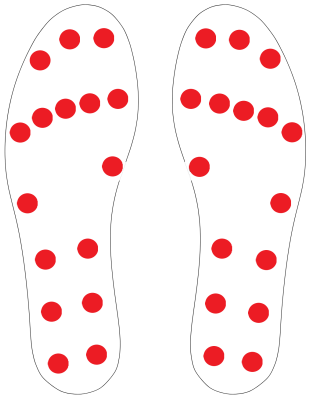

Figure 5. Layout of the individual pressure sensors in the Moticon OpenGo instrumented insoles. Each dot represents a sensor.

6.1. Multi-Stream Spatiotemporal Encoder

To exploit the heterogeneity of the input modalities, we adopt a multi-stream architecture composed of three dedicated encoders: a pressure encoder based on a Graph Convolutional Network (GCN), an IMU/force encoder based on an LSTM–MLP stack, and a pose encoder used only during training.

Pressure Encoder. The plantar pressure vector $\mathbf{P}_t$ is decomposed into left/right foot grids, each represented as an independent undirected graph whose nodes correspond to pressure sensors. A spatial GCN is applied to each foot graph:

$$\mathbf{H}^{(k+1)} = \sigma\left(\widetilde{\mathbf{D}}^{-\frac{1}{2}}\widetilde{\mathbf{A}}\widetilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(k)}\mathbf{W}^{(k)}\right),$$

where $\widetilde{\mathbf{A}}$ is the sensor adjacency matrix, $\widetilde{\mathbf{D}}$ its degree matrix, and σ a ReLU activation. The resulting features are concatenated and passed to an LSTM capturing temporal pressure evolution. LSTM layers do not reduce the dimensionality of their inputs.

IMU/Force Encoder. The auxiliary vector $\mathbf{A}_t$ is processed by an LSTM extracting global foot dynamics, followed by a lightweight MLP. This branch provides complementary information describing acceleration profiles, foot orientation, and load distribution.

Pose Encoder (Training Only). During training, a third encoder processes the ground-truth pose sequence $\mathbf{J}_{1:T}$ to produce a latent representation $\mathbf{z}^{\text{pose}}_{1:T} \in \mathbb{R}^{576}$. Its architecture mirrors the main model and includes a GCN operating on a kinematic graph of the lower body (without arms), followed by an LSTM. This encoder is disabled at inference.

6.2. Two-Stream Transformer Fusion

The outputs of the pressure and IMU encoders are temporally aligned and concatenated before entering a Two-Stream Transformer (see Figure 7). For each layer containing self-attention and cross-attention blocks, the Transformer produces a fused representation

$$\mathbf{F}_{1:T} \in \mathbb{R}^{T\times 576}.$$

The dimensionality of 576 is fixed throughout the Transformer and matches that of the latent pose representation used for the feature consistency loss.

6.3. Latent Projection and Motion Decoder

The fused features are projected through an MLP into a compact latent code

$$\mathbf{z}_{1:T} \in \mathbb{R}^{576},$$

followed by a final MLP decoder mapping the latent sequence to 3D joint coordinates:

$$\hat{\mathbf{J}}_{1:T} = g(\mathbf{z}_{1:T}) \in \mathbb{R}^{T\times 72}.$$

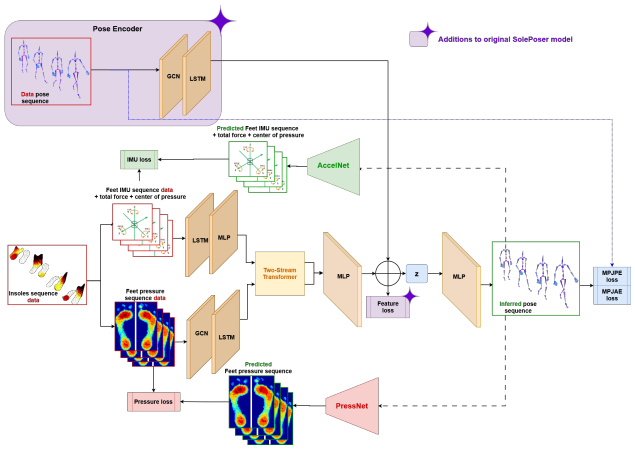

Figure 6. Baseline model for 3D pose inference.

6.4. Cycle-Consistency via PressNet and AccelNet

Following SolePoser [26], we integrate two pretrained networks: PressNet (see Figure 9), which reconstructs plantar pressure maps from predicted poses, and AccelNet (see Figure 8), which reconstructs insole acceleration data. Given the predicted pose $\hat{\mathbf{J}}_{1:T}$, these networks produce

$$\hat{\mathbf{P}}_{1:T}, \qquad \hat{\mathbf{A}}^{\text{acc}}_{1:T}.$$

The reconstruction losses

$$\mathcal{L}_{\text{press}} = \|\hat{\mathbf{P}}_{1:T} - \mathbf{P}_{1:T}\|_1, \qquad \mathcal{L}_{\text{acc}} = \|\hat{\mathbf{A}}^{\text{acc}}_{1:T} - \mathbf{A}^{\text{acc}}_{1:T}\|_1,$$

serve as cycle-consistency constraints, enforcing biomechanical plausibility of the predicted poses.

6.5. Feature Consistency Loss

Because both the insole-based pipeline and the pose encoder produce latent features of identical dimensionality, we enforce feature-level alignment through

$$\mathcal{L}_{\text{feat}} = \|\mathbf{z}_{1:T} - \mathbf{z}^{\text{pose}}_{1:T}\|_2^2.$$

This constraint encourages the model to learn a representation that is simultaneously predictive from insole data and self-consistent with pose-derived latents.

6.6. Training and Inference

The total training loss is a weighted sum where MPJPE and MPJAE supervise absolute joint positions and accelerations, respectively. During inference, only $(\mathbf{P}_{1:T}, \mathbf{A}_{1:T})$ are provided. The pose encoder and cycle modules are disabled, and the model outputs $\hat{\mathbf{J}}_{1:T}$ directly through the insole-based pipeline.

6.7. Results on different body parts

Table 4 reports per-joint errors, as well as aggregated results over upper- and lower-body segments. A pronounced

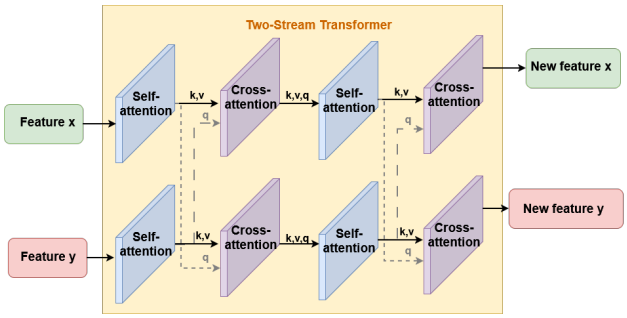

Figure 7. Two-Stream Transformer, as proposed by [26].

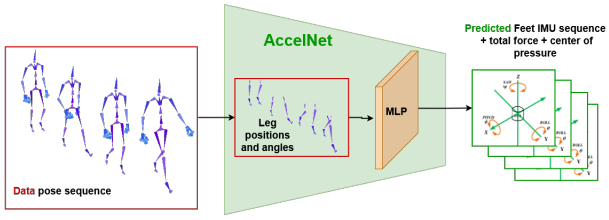

Figure 8. AccelNet network, as proposed by [26].

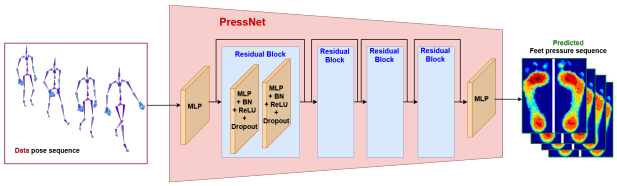

Figure 9. PressNet network, as proposed by [26].

performance gap is observed: upper-body joints (restricted here to the trunk, neck, and head) are estimated substantially more accurately than lower-body joints. This trend, although counter-intuitive given that the model relies exclusively on foot-mounted sensors, can be attributed to structural factors. As detailed in section 3.1, all poses are expressed in a pelvis-centered coordinate frame, which inherently favors joints proximate to the pelvis while amplifying the apparent motion of distal joints. Combined with the greater kinematic variability of the legs and feet relative to the trunk, this representation likely accounts for the lower-body degradation observed in the results.

6.8. Ablation studies

Several ablation studies were realized on the model illustrated by Figure 6, studying the impact of various architecture components and loss functions on its performance.

6.8.1. MPJAE loss

The MPJAE and MPJPE are both used as loss functions during training, allowing the model to focus on the position of the joints and the angles between them. A first ablation

Body part	MPJPE ↓ (cm)	Instability ↓ (cm)	MPJAE ↓ (°)
Head	41.2 ± 1.9	10.9 ± 0.5	–
Neck	29.6 ± 1.3	8.6 ± 0.3	13.9 ± 1.7
Spine 2	20.7 ± 0.8	5.9 ± 0.2	4.5 ± 0.8
Left Shoulder	27.7 ± 1.6	7.7 ± 0.3	–
Right Shoulder	27.6 ± 1.5	7.7 ± 0.3	–
Spine 1	12.3 ± 0.3	3.5 ± 0.1	4.6 ± 0.7
Spine	5.5 ± 0.2	1.9 ± 0.03	7.2 ± 0.3
Hips	5e-7 ± 5e-7	8e-7 ± 9e-7	11.7 ± 2.9
Upper Body	20.6 ± 1	5.8 ± 0.2	6.2 ± 0.8
Left Up Leg	7.1 ± 0.1	2.7 ± 0.1	11.9 ± 1.4
Left Leg	34.5 ± 2.7	9.5 ± 0.7	35.4 ± 3.2
Left Foot	57.4 ± 2.6	11.4 ± 0.4	27.7 ± 6.9
Left Toe Base	67.6 ± 3.5	13.2 ± 0.7	–
Right Up Leg	7.1 ± 0.1	2.6 ± 0.1	13.2 ± 1.2
Right Leg	34 ± 2.9	9.5 ± 0.5	35.2 ± 4.3
Right Foot	58.4 ± 3.3	11.3 ± 0.3	26.1 ± 3.9
Right Toe Base	67.7 ± 3.5	13.2 ± 0.5	–
Lower Body	41.7 ± 2.3	9.2 ± 0.4	20.5 ± 3.3
Global	31.1 ± 0.8	7.5 ± 0.1	13.3 ± 1.5

Table 4. Results on pose inference on each joint.

MPJAE loss	MPJPE ↓ (cm)	Instability ↓ (cm)	MPJAE ↓ (°)
w/	31.1 ± 0.8	7.5 ± 0.1	13.3 ± 1.5
w/o	33.9 ± 3.5	7.3 ± 0.8	41.7 ± 18.4

Table 5. Results w/ and w/o the use of MPJAE as training loss.

Insoles reconstruction	MPJPE ↓ (cm)	Instability ↓ (cm)	MPJAE ↓ (°)
Yes	31.1 ± 0.8	7.5 ± 0.1	13.3 ± 1.5
No	26 ± 0.6	7.2 ± 0.1	13.2 ± 1.5

Table 6. Results w/ and w/o the use of AccelNet and PressNet.

study is proposed to understand the impact of using the MPJAE as a loss function on the model's performance. Table 5 shows the results of this ablation study.

Unsurprisingly, using the MPJAE as a loss function leads to much lower MPJAE during the validation phase. However it also leads to lower MPJPE, showing the interest of using this loss function even when the MPJPE is the sole indicator of performance. Moreover, not using this loss function leads to more unstable results, as shown by the higher standard deviations in metric values obtained in this case. It can be concluded that the MPJAE is a generally valuable information to provide to the model during training for a pose inference task.

6.8.2. AccelNet and PressNet

During training, the inferred pose sequence is passed through AccelNet and PressNet to reconstruct, respectively, insole accelerations and plantar pressure. The resulting reconstruction errors define an auxiliary loss term used to supervise pose prediction. To assess the influence of these networks and of the associated reconstruction loss, we conduct an ablation study; the results are reported in Table 6.

Contrary to the recommendation of [26], incorporating

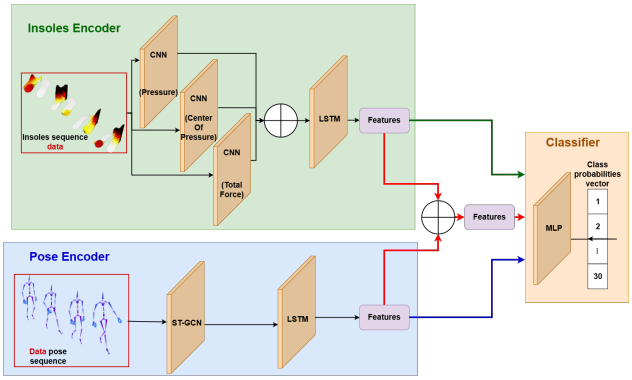

Figure 10. Action-classification framework.

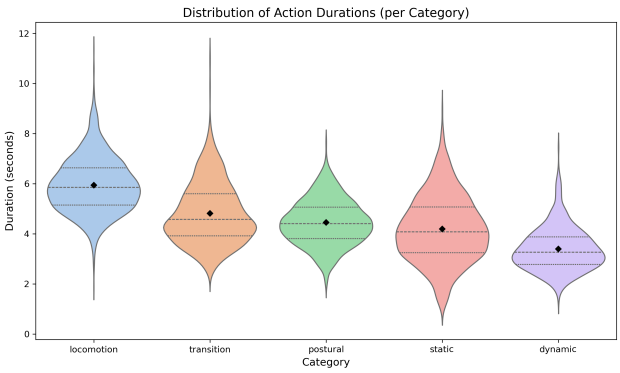

Figure 11. Action durations per category across the entire dataset.

AccelNet and PressNet into our framework degrades performance. We attribute this effect to the central architectural difference between our model and SolePoser: the presence of a dedicated pose encoder. SolePoser (which uses AccelNet/PressNet but no pose encoder) and our variant without AccelNet/PressNet (but with a pose encoder) both outperform our full model combining all components. These findings suggest that, although the pose encoder and the cycle-consistency networks independently provide benefits, integrating them effectively is non-trivial. A different strategy appears necessary to leverage their complementary strengths.

7. Insole-based activity Recognition

Our framework relies on two encoders (Figure 10). The insole encoder maps each frame to three compact descriptors extracted by independent CNN branches: a pressure vector $\mathbf{p} \in \mathbb{R}^{32\times1}$, a center-of-pressure vector $\mathbf{c} \in \mathbb{R}^{4\times1}$, and a total-force vector $\mathbf{f} \in \mathbb{R}^{2\times1}$. These vectors are concatenated and processed through a uni-directional LSTM over a 3-second temporal window ($T = 300$), producing a modality-specific latent representation.

The pose encoder operates on MoCap data consisting of

24 joints coordinates. The sequence is first processed by a three-layer spatial–temporal GCN, followed by an LSTM over a 3-second window ($T = 360$), yielding the pose-based feature representation. A shared MLP classifier then takes as input either one or both latent representations, depending on the training strategy.

The 3-second temporal window was based on an analysis of action durations across the entire dataset (Figure 11). For each movement, the analysis starts at the beginning of the action and considers the first 3 seconds. For examples shorter than this window, we apply a thresholding strategy: if more than 60% of the action duration belongs to a single class, the example is assigned the majority class.

7.1. Training Strategies

We evaluate three training configurations (see Figure 10) to assess the contribution of each modality and their combination. In the *insoles-only* setup, the classifier is trained and tested exclusively on features extracted from the insole encoder. Conversely, the *MoCap-only* configuration relies solely on pose-based features for both training and evaluation. In the *concatenated multimodal* setup, features from both encoders are concatenated before being fed into the classifier, allowing the joint exploitation of plantar-pressure and kinematic cues.

Finally, We explore a a student–teacher paradigm for a realistic use case: at inference time, only the insole encoder is required, enabling deployment without MoCap data while still benefiting from the knowledge transferred from the teacher. In this approach, the classifier is trained using both modalities, with the pose encoder acting as a teacher to guide the insole encoder through feature-alignment losses.

As reported in Table 7, the concatenation of MoCap and insole features consistently achieves the best performance in the majority of cases. The multimodal fusion provides a richer representation of the movement, which is particularly useful for distinguishing between similar actions. Leveraging multiple modalities allows the model to capture complementary aspects of motion, which is critical for nuanced classification of closely related movements. Table 8 presents the same analysis on a per-subject basis, confirming that the benefits of multimodal fusion generalize across individual subjects.

When using insole features alone, classification is highly effective for locomotion-related actions, where foot pressure patterns are strongly discriminative. However, performance decreases for actions primarily involving the upper body or when there is no ground contact, highlighting the limitations of plantar-pressure information in isolation. Analyzing complex movements with only insole data remains a challenging task, which has been relatively underexplored in the literature and existing datasets. In this context, the student–teacher paradigm offers a compelling solution: by distilling knowledge from the MoCap encoder into the insole encoder during training, the model can leverage rich kinematic information while relying solely on insoles at inference. This approach improves performance and provides a promising avenue for applications where vision-based systems are impractical, although action classification from insole data alone remains a highly complex problem.

7.2. Student-Teacher Loss

We evaluated several loss functions for the student–teacher strategy, following a distillation-based approach in which the teacher guides the student. The overall loss $\mathcal{L}$ is defined as a weighted combination of the standard classification loss $\mathcal{L}_{\text{cls}}$ and the distillation loss $\mathcal{L}_{\text{distill}}$:

$$\mathcal{L} = \alpha\, \mathcal{L}_{\text{distill}} + (1 - \alpha)\, \mathcal{L}_{\text{cls}}, \tag{8}$$

where α balances the contribution of the teacher's guidance and the ground-truth supervision. We varied α from 25% to 75% and found that setting $\alpha = 0.5$ yields the best overall performance, providing an effective trade-off between learning from the teacher and from the ground-truth labels.

	ClassName	Insoles (%)	Skeleton (%)	Insoles+Skeleton (%)	Insole/Skeleton (%)
Static	Standing upright without moving	42.03 ± 21.11	65.05 ± 27.47	69.22 ± 22.67	46.86 ± 26.38
	Balancing on the left leg	71.83 ± 25.08	74.31 ± 22.52	71.17 ± 22.34	73.12 ± 23.19
	Balancing on the right leg	71.01 ± 24.71	69.38 ± 17.35	71.68 ± 21.65	65.75 ± 22.39
	Doing squats	70.11 ± 23.47	88.62 ± 14.15	87.37 ± 17.64	72.13 ± 25.60
	Standing on tiptoes	76.79 ± 17.64	72.54 ± 17.01	83.08 ± 15.75	77.44 ± 15.76
	Standing on heels	79.51 ± 21.19	76.35 ± 18.71	79.86 ± 19.02	77.57 ± 23.03
Locomotion	Walking	89.90 ± 12.42	90.60 ± 11.47	94.90 ± 6.75	86.37 ± 17.01
	Jogging	92.15 ± 13.62	92.15 ± 11.29	95.28 ± 8.05	91.22 ± 12.89
	Walking backwards	82.78 ± 15.69	87.78 ± 12.63	89.97 ± 12.18	82.47 ± 15.64
	Walking on tiptoes	89.90 ± 20.04	81.67 ± 24.51	92.08 ± 15.67	87.42 ± 17.12
	Hopping around on one leg	78.91 ± 17.69	76.13 ± 18.92	78.92 ± 21.45	75.64 ± 21.17
	Side stepping	77.43 ± 25.56	87.40 ± 20.65	86.43 ± 17.58	76.02 ± 24.63
Dynamic	Jumping in place	70.38 ± 28.54	92.26 ± 13.87	88.75 ± 15.75	70.34 ± 29.62
	Jumping on the right leg	76.67 ± 16.92	77.00 ± 25.03	77.19 ± 21.19	74.50 ± 20.48
	Jumping on the left leg	50.99 ± 25.79	61.00 ± 27.97	55.01 ± 25.42	52.41 ± 27.20
	Jumping to the side	78.60 ± 21.37	85.96 ± 17.52	83.11 ± 17.74	63.84 ± 26.14
	Jumping forward	79.84 ± 23.61	93.40 ± 9.38	91.28 ± 13.69	77.03 ± 31.44
	Jumping backward	75.10 ± 25.53	90.71 ± 14.66	87.20 ± 19.80	74.34 ± 24.28
Interaction	Sitting on the stool	97.15 ± 6.39	98.40 ± 4.55	97.78 ± 4.97	98.09 ± 4.04
	Climbing the stairs	94.38 ± 7.59	97.50 ± 9.16	96.84 ± 6.49	91.25 ± 12.12
	Lifting the bag with one hand	42.48 ± 20.83	80.72 ± 23.15	78.93 ± 20.03	43.15 ± 26.51
	Lifting the bag with two hands	61.55 ± 20.67	76.41 ± 24.95	80.00 ± 19.08	54.23 ± 28.33
	Pulling the bag	58.86 ± 30.82	65.60 ± 32.86	71.34 ± 30.84	58.12 ± 32.99
	Pushing the bag	60.52 ± 25.50	72.57 ± 27.85	72.42 ± 23.36	47.90 ± 28.05
Postural	Leaning forward	39.48 ± 24.27	79.42 ± 22.95	74.51 ± 24.54	50.93 ± 25.77
	Leaning backward	52.06 ± 23.26	93.06 ± 11.22	87.46 ± 17.96	56.28 ± 22.41
	Leaning to the side	56.53 ± 20.61	87.33 ± 22.90	90.83 ± 17.53	67.33 ± 21.64
	Looking backward	72.19 ± 20.87	84.41 ± 24.25	85.07 ± 19.54	71.08 ± 25.99
	Putting left hand on the ground	36.56 ± 22.23	83.44 ± 19.61	72.50 ± 26.88	32.50 ± 24.49
	Putting right hand on the ground	41.41 ± 20.92	77.66 ± 24.20	78.28 ± 19.20	36.72 ± 26.26

Table 7. Class-wise performance grouped by categories (LOSO).

	Insoles (%)					Insoles/Mocap (%)				
	Static	Locomotion	Dynamic	Interaction	Postural	Static	Locomotion	Dynamic	Interaction	Postural
S01	78.69	93.33	71.67	68.33	96.67	75.41	90.00	85.00	75.00	93.33
S02	93.10	96.67	88.33	82.26	96.67	91.38	91.67	70.00	87.10	95.00
S03	90.16	83.33	78.33	77.97	88.33	85.25	96.67	73.33	74.58	81.67
S04	86.67	98.33	61.67	86.67	98.33	86.67	95.00	58.33	88.33	95.00
S05	91.67	85.00	78.33	77.05	90.00	95.00	86.67	88.33	72.13	95.00
S06	81.97	91.67	76.67	57.63	90.00	81.97	73.33	95.00	69.49	88.33
S07	96.61	86.89	90.32	77.42	75.00	98.31	86.89	100.00	51.61	70.00
S08	81.67	96.67	81.67	65.00	100.00	83.33	95.00	95.00	58.33	100.00
S09	90.16	86.67	75.00	81.36	96.67	90.16	86.67	58.33	71.19	98.33
S10	80.65	86.67	73.33	74.14	93.33	82.26	85.00	75.00	94.83	81.67
S11	88.33	90.00	93.33	73.33	86.67	90.00	93.33	56.67	83.33	76.67
S12	89.47	96.83	83.33	81.67	71.67	94.74	84.13	78.33	78.33	51.67
S13	84.31	87.04	69.09	69.64	92.59	86.27	85.19	92.73	69.64	96.30
S14	85.25	88.33	79.66	81.67	83.33	86.89	83.33	76.27	85.00	98.33
S15	85.00	91.67	83.33	60.00	86.67	88.33	93.33	93.33	45.00	85.00
S16	90.16	85.00	90.16	71.19	80.00	86.89	88.33	93.44	66.10	83.33
S17	86.89	90.00	68.33	76.67	91.67	80.33	83.33	81.67	58.33	85.00
S18	96.67	75.00	75.00	88.33	90.00	96.67	80.00	88.33	80.00	78.33
S19	96.67	86.67	81.67	61.67	100.00	91.67	80.00	86.67	55.00	100.00
S20	91.67	83.33	91.53	79.66	93.33	88.33	73.33	91.53	64.41	95.00
S21	75.00	100.00	98.33	56.67	95.00	83.33	90.00	96.67	55.00	91.67
S22	95.00	96.67	86.67	83.05	81.67	78.33	98.33	80.00	77.97	85.00
S23	78.33	95.00	71.67	68.33	96.67	81.67	88.33	78.33	65.00	85.00
S24	81.97	90.00	66.67	76.27	93.33	83.61	88.33	91.67	88.14	100.00
S25	93.22	96.67	83.05	82.26	96.67	94.92	98.33	81.36	85.48	96.67
S26	90.16	90.16	81.67	73.33	95.00	96.72	88.52	81.67	75.00	90.00
S27	64.91	91.94	61.67	77.97	90.00	71.93	91.94	60.00	71.19	80.00
S28	79.66	96.77	70.00	74.55	95.00	88.14	91.94	70.00	74.55	90.00
S29	86.67	90.00	64.41	73.33	71.67	95.00	90.00	66.10	71.67	83.33
S30	85.96	71.67	57.63	83.05	98.33	80.70	73.33	81.36	79.66	80.00
S31	93.44	85.00	73.33	75.00	91.67	91.80	80.00	90.00	76.67	80.00
S32	79.66	96.72	68.33	90.00	91.67	84.75	98.36	81.67	85.00	90.00
	86.55	89.99	77.32	75.17	90.55	87.21	87.77	81.13	72.91	87.49
	± 6.99	± 6.40	± 9.94	± 8.36	± 7.52	± 6.37	± 6.90	± 11.94	± 11.73	± 10.08

Table 8. Subject-wise performance (LOSO).