Knowledge Distillation from a Raw-Signal Foundation Model to a Feature-Based On-Device Sleep Stager

Lucas Fontes Buzuti[1], Daniel Ide Futata[1], Jhonatas Conceição[1], Felipe Chaud Pinheiro[1], Jean Pieri Angelo de Matos[1], Byeong-ho Lee[2], Seungman Yang[2], William Camilo Ariza-Zambrano[1]

[1] Samsung R&D Institute Brazil [2] Health AI Algorithm Lab, Samsung Electronics

NeurIPS 2026 Workshop · On-Device Intelligence: Foundation Models under Real-World Constraints

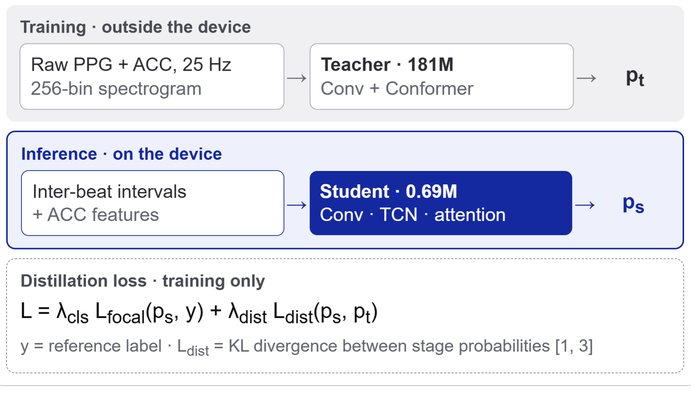
Problem

- The wearable device cannot run the large teacher model
- The deployed classifier uses derived features
- The student has the same inputs and architecture as the deployed classifier, but the teacher has different inputs and architecture

Method

- **Teacher:** PPG + ACC spectrograms · 181M parameters · HuBERT-style pretraining + sleep-stage fine-tuning [1, 2]
- **Student:** inter-beat intervals + ACC features [4] · 0.69M parameters · TCN
- Both models use aligned data from the same recording

Figure A · Two input paths from the same overnight recording



The teacher and student use different inputs from the same recording. Only the student and its feature pipeline operate on the device.

Data and training

- 19 studies · nine wearable models
- Participant-level split: 2,764 train / 339 validation / 575 test
- Four classes: wake, light (N1/N2), deep (N3), REM
- Loss weights: $\lambda_{cls} = 0.75$, $\lambda_{dist} = 0.25$
- Teacher updates start after 20,000 steps

Model	Input and architecture
Teacher	PPG + ACC spectrogram; convolutional layers + Conformer
Student	Inter-beat intervals + ACC features; convolutional front end, time encoding, TCN, and attention

72.11%
Student macro F1

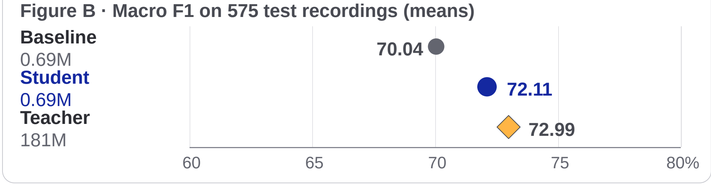
+2.07
points vs. baseline, same architecture

≈0.38%
of teacher parameters (0.69M / 181M)

Primary results

- Macro F1: **72.11%** student vs. 70.04% baseline vs. 72.99% teacher
- Balanced accuracy: 78.48% → 78.63%
- Same architecture and parameter count as the baseline

Figure B · Macro F1 on 575 test recordings (means)



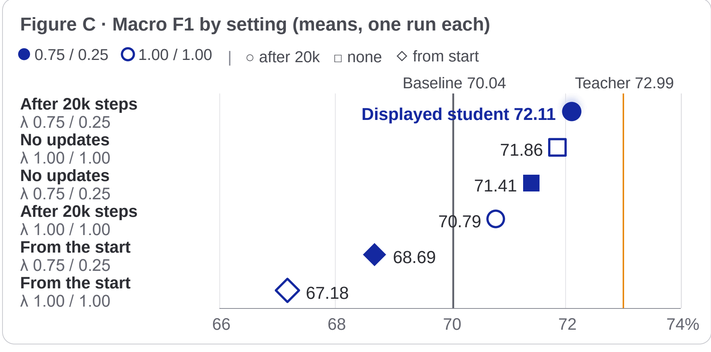
Means across recordings; SD does not show uncertainty in model differences. Comparisons are descriptive.

Model	Params	Macro F1 (%)	Cohen's κ	Bal. acc. (%)
Baseline	0.69M	70.04 ± 12.66	0.63 ± 0.14	78.48 ± 10.54
Teacher	181M	72.99 ± 11.87	0.67 ± 0.13	80.11 ± 9.19
Student	**0.69M**	**72.11 ± 12.58**	**0.66 ± 0.14**	**78.63 ± 10.51**

Teacher update schedule

- Three update schedules × two loss settings
- Delayed or no updates > updates from the start

Figure C · Macro F1 by setting (means, one run each)

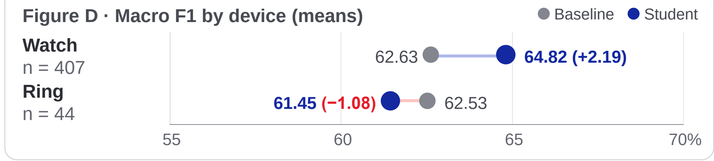


Each setting has one training run. The results do not identify the best teacher update schedule.

Results on a separate corpus

- 407 watch + 44 ring PSG-labelled recordings, unseen in training
- Watch macro F1: **+2.19 points**
- Ring macro F1: −1.08 points (no ring training data)

Figure D · Macro F1 by device (means)



Balanced accuracy: watch +0.97 points; ring −0.41 points.

Stage results

Metric (%)	Watch, n = 407	Ring, n = 44
Light-sleep sensitivity	75.09 → 81.10 **+6.01**	79.83 → 81.68 **+1.85**
Deep-sleep sensitivity	64.48 → 69.92 **+5.44**	60.11 → 70.99 **+10.88**
REM-sleep sensitivity	82.02 → 84.52 **+2.50**	86.26 → 89.09 **+2.83**
Wake precision	47.27 → 62.11 **+14.84**	51.10 → 65.15 **+14.05**

Baseline → student means; differences in points.

Discussion

- **Near the teacher:** 72.11% vs. 72.99% macro F1 at ≈0.38% of its parameters
- **No deployment change:** same inputs, architecture, and 0.69M parameters
- **Holds on unseen data:** +2.19 macro F1 on the watch corpus
- **Large stage gains:** wake precision +14.84 / +14.05; ring deep-sleep sensitivity +10.88

Limitations: wake sensitivity −9.30 / −13.27; ring deep-sleep precision 31.22% → 28.20%; one task; no significance or efficiency tests.

Conclusion

- Higher macro F1 on the first corpus and separate watch corpus, with the existing inputs and architecture
- Ring corpus: lower macro F1 with distillation

References

[1] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network. 2015. arXiv:1503.02531.
[2] Hsu W-N et al. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. 2021. arXiv:2106.07447.
[3] Lin T-Y et al. Focal Loss for Dense Object Detection. 2017. arXiv:1708.02002.
[4] Topalidis P et al. The Virtual Sleep Lab—A Novel Method for Accurate Four-Class Sleep Staging Using Heart-Rate Variability from Low-Cost Wearables. Sensors. 2023;23(5):2390.

We thank Orlem Santos, Hyojin Lee, and Eugene Hwang for discussions about this work.