Text2IMU: Advancing Human Activity Recognition by Text-Driven IMU Data Synthesis

1st Lars Ole Haeusler
Intelligent Embedded Systems Lab
University of Freiburg
Freiburg, Germany
haeusler@informatik.uni-freiburg.de

2nd Lena Uhlenberg
Intelligent Embedded Systems Lab
University of Freiburg
Freiburg, Germany
uhlenberg@informatik.uni-freiburg.de

3rd Oliver Amft
Intelligent Embedded Systems Lab
University of Freiburg, Hahn-Schickard
Freiburg, Germany
amft@ieee.org

***Abstract*—Inspired by the progress of motion synthesis models, we leverage cross-modality transfer to generate realistic synthetic Inertial Measurement Unit (IMU) data from textual descriptions, hence Text2IMU. We use an established motion synthesis model and textual descriptions to generate sequences of 3D human activities. To obtain realistic and diverse sensor readings, we created multiple body surface models with different body morphologies. With the text prompts, we let the surface models perform activities and synthesise acceleration and gyroscope data for multiple virtual IMU positions. We show that synthetic data, generated by Text2IMU, can be used to classify activities across three public benchmark datasets. We demonstrate that our Text2IMU synthesis approach does not require measured data of the target domain. Text2IMU yields an average Human Activity Recognition (HAR) accuracy of 79.2% for correctly synthesised activities, which doubles the performance of synthetic sensor data obtained from baseline models. We demonstrate that synthetic HAR model training can replace empirical data acquisition when the prompted activities can be successfully generated.**

***Index Terms*—Synthetic Data, Human Activity Recognition, IMU sensor synthesis, Generative Motion Synthesis**

I. Introduction and related work

HAR is used to estimate activities across various applications, e.g., human behaviour understanding, safety, and security monitoring, as well as tracking of daily routines [1], [2]. A major limitation of current HAR systems is the lack of adequate amounts of training data. Expert labour required for preparation, recording, and labelling of sensor data often limits the size of IMU datasets. To address the prevalent data scarcity, a current trend in HAR research focuses on the use of synthetic data in the training process. Especially, cross-modality methods became dominant in recent years to synthesise IMU sensor readings. For example, online videos could be used to track and translate human movements into corresponding IMU data [3]. Another approach utilised personalised biomechanical dynamics models and human surface models based on Motion Capture data to synthesise IMU sensor time series data [4].

Human motion synthesis, based on textual descriptions, can also be employed to generate synthetic IMU data. Large datasets, e.g., HumanML3D [5], include various motion sequences in the form of joint rotations with corresponding textual descriptions of the respective activity. HumanML3D was used to train text-based human motion synthesis models, e.g. T2M-GPT [6]. Leng et al. [7], [8] utilised T2M-GPT and proposed the IMUGPT framework to generate synthetic accelerometer data. The authors used ChatGPT to create textual prompts of activities, which were then fed into T2M-GPT. The generated joint representations were used to synthesise acceleration data according to their motion trajectories. Ray et al. [9] followed a similar approach in the Text-to-Pressure framework. Compared to IMUGPT, the authors used Joints2SMPL [10] to translate the joint representations into Skinned Multi-Person Linear (SMPL) models [11]. Surface models were subsequently used to generate synthetic pressure maps, which were utilised to classify human activities.

One common drawback of current synthetic IMU data synthesis methods is the remaining domain gap between synthesised data and measured data. As a result, HAR performance when trained solely on synthetic data cannot compete with measured data [3], [12]. Approaches to mitigate the domain gap require either measured data from the target dataset to adapt the synthetic data in a calibration step [3], [7] or mixing of synthetic data with measured data during training [4]. In conclusion, previous approaches require measured sensor data and thus manual data collection for every single domain/application prior to deployment.

In this work, we propose a framework that generates realistic synthetic IMU data based on textual activity descriptions. The framework can be used to classify human activities, without requiring any measured data of the target domain, thus serves as a proof of principle for future HAR application development without prior data collection. As a minimum, our approach requires the following information only: (1) prompt of the activities to be recognised, and (2) IMU sensor positions. Compared to the previous works, e.g., IMUGPT or Text-to-Pressure, we generate multiple, diverse SMPL surface models resembling different body morphologies for each motion representation. Furthermore, we simulate multiple, varying virtual IMU sensor placements to synthesise accelerometer and gyroscope data. Finally, we add multiple data augmentation strategies to simulate different sensor orientations and noise.

This paper provides the following contributions:

1) We synthesise acceleration and gyroscope data from textual descriptions and human surface models with varying body morphologies. We vary virtual sensor positions and

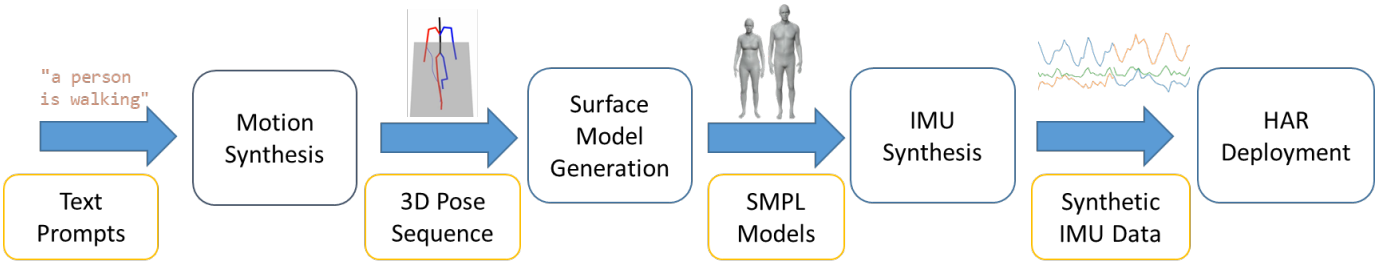

Fig. 1. Overview of the Text2IMU framework. Textual descriptions are used to generate human activities as a 3D pose sequence. Pose sequences are converted into multiple surface models that represent humans with different body morphologies. For a nominal body position of each real IMU, multiple virtual sensor placements are chosen to synthesise accelerometer and gyroscope data. After further augmentation, the synthetic data is used for HAR model training.

employ data augmentation to bridge the domain gap between the virtual and real IMU data.

2) We synthesise IMU data according to activities and body positions described in three public benchmark datasets. We train HAR models exclusively on non-calibrated, synthetic IMU data and evaluate HAR model performance against classically trained models using measured IMU datasets. We show that HAR performance based on synthetic data training can reach the performance of HAR models trained from measured data when the activities can be meaningfully generated.

II. Methods

An overview of the Text2IMU framework is shown in Fig. 1.

A. Human motion synthesis

We used ChatGPT to generate 10 different textual descriptions for each activity by providing a general structure, length constraints, and few examples. Execution variations depend on, e.g., speed and location of the person. For example, two generated prompts for walking are: "The person is walking on a track, following the oval path.", "The person is walking slowly with multiple strides.". We utilised T2M-GPT [6] to synthesise human motion from the textual descriptions. The obtained motion representation consists of a sequence of 22 joint positions in a 3D space with 20 samples/s, typically varying from 3-10 s.

B. Surface model generation

To accurately simulate virtual sensor placements, the obtained motion representation (stick figure in Fig. 1) was converted to a SMPL surface model with joints2smpl [10]. The SMPL model generates a triangulated mesh of pose parameters θ and shape parameters β. To simulate different participants performing the desired activities, we use a set of varying shape parameters β [13] to obtain 8 different SMPL models (4 male, 4 female) performing the same activities. Body proportions were chosen to represent the tallest and shortest model for each gender ($|\beta_i| < 5, \quad$ for all $i = 1, 2, \ldots, 10$), one model with average body measurements[1], and one model with average height but an adipose Body-Mass-Index of 35 kg/m^2.

[1]https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Gesundheitszustand-Relevantes-Verhalten/Tabellen/liste-koerpermasse.html

C. IMU synthesis

To synthesise IMU data and represent IMU placement variations, we selected nine vertices of the SMPL model surface as sensor positions according to the measured IMU position used in benchmark datasets. To derive acceleration and angular velocity sensor readings, we followed a similar approach as proposed by Uhlenberg et al. [4], [14]. Each vertices can be described as a position vector $\vec{k}(t) = (k_x(t), k_y(t), k_z(t))$ for any time t in a coordinate frame $G \in \mathbb{R}^{3x3}$. The linear acceleration of each vertex $\vec{a}(t) = (a_x, a_y, a_z)$ was calculated as sum of the dynamic sensor acceleration $\vec{a}_d(t)$ and the equivalent gravitational acceleration $\vec{a}_g(t)$ exerted on the sensor device, expressed as: $\vec{a}(t) = \vec{a}_d(t) + \vec{a}_g(t)$. Dynamic sensor acceleration $\vec{a}_d(t)$ was derived as second derivative of vector $\vec{k(t)}$ according to $\vec{a_d}(t) = \frac{d^2\vec{k}(t)}{dt^2}$. Gravitational acceleration $\vec{a}_g(t)$ was obtained by multiplying the unit vector $\vec{k}_j(t)$, which describes the gravity contribution along each axis of the sensor, with the gravity constant $g = 9.81\,\text{m/s}^2$ according to $\vec{a}_g(t) = \vec{k}_j(t) \cdot g$. To obtain $\vec{k}_j(t)$, the gravity-carrying axis $\vec{j}$ was determined and multiplied with the inverse of the rotation matrix Q_S^G, which represents the sensor's global orientation, according to $\vec{k}_j(t) = (Q_S^G)^{-1} \cdot \vec{j}$. The angular velocity was subsequently determined from quaternion orientation estimates q_S^G of the rotation matrix Q_S^G as $\vec{\omega}(t) = \frac{q_S^G \vec{k}(t)}{dt}$. An example of synthesised and measured data is depicted in Fig. 2.

D. Synthetic activity quality assessment

To evaluate synthetic activity quality achieved by T2M-GPT, five independent raters (blinded to the HAR results) reviewed video clips of each generated activity and rated how well the activity matched the prompt on a 5-point Likert Scale [15].

Based on the average ratings of the generated activities, we defined three activity class subsets per benchmark dataset: Activity Set 1 included four activities with an average rating above 75%: *Walking, (Rope) Jumping, Running*, and *Sitting*. Activity Set 2 included six activities with a minimum rating of 50%, thus covering Set 1 plus *Stairs Down* and *Lying*. Activity Set 3 consists of all activities that are available in each dataset. The WISDM dataset does not include *Jumping* and *Lying* activities, thus Set 1 comprised three activities and Set 2 four activities.

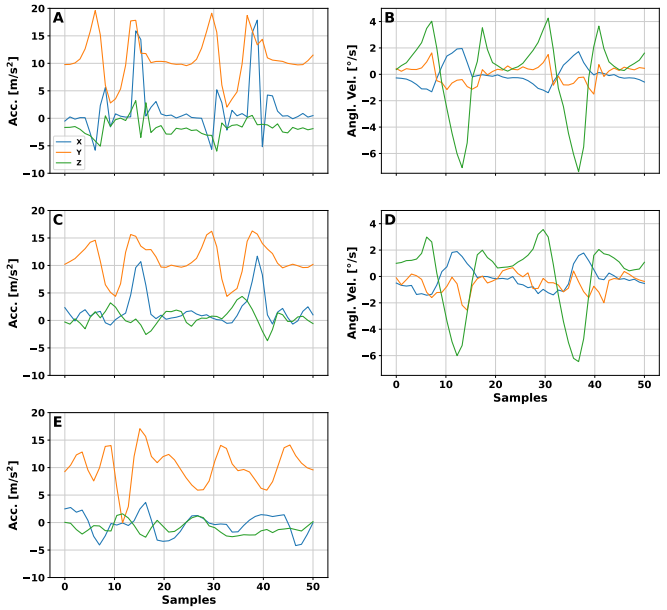

Fig. 2. Comparison of IMU data based on example data at the shank during walking (RealWorld dataset). A: Measured accelerometer data. B: Measured gyroscope data. C: Text2IMU accelerometer data. D: Text2IMU gyroscope data. E: IMUGPT⋆ accelerometer data. No gyroscope data available from IMUGPT. IMUGPT⋆ was used without synthetic data calibration (i.e., without adapting by measured data) as fair comparison to our approach.

E. HAR evaluation

We evaluated our Text2IMU synthesis approach with three frequently used public benchmark datasets: RealWorld [16], PAMAP2 [17], and WISDM [18]. For each benchmark dataset, we generated synthetic IMU data using simulated sensors positioned according to the placement descriptions provided by the respective datasets. Measured IMU data of the benchmark datasets were downsampled to 20 Hz to match synthetic data.

To improve generalisation of synthetic data to sensor rotations in target datasets, we generated six possible permutations of the virtual IMU sensor axes (x, y, z) following [12]. Additionally, we add Gaussian noise with a zero mean to the synthetic data to imitate real sensor noise. The total synthetic dataset size was 64 h 13 min for Realworld, 106 h 21 min for PAMAP2, and 46 h 39 min for WISDM.

For classification, we employed DeepConvLSTM [19], a model widely used in HAR research [3], [4], [7]. Input data was divided into 2 s segments corresponding to 40 samples each, using a sliding window with 50% overlap. We provide two baselines for each experiment. Firstly, we performed a leave-one-participant-out (LOPO) cross-validation based on the measured data of each benchmark dataset. As second baseline, we used IMU data synthesised by IMUGPT [7]. We generated IMUGPT⋆ data without calibration (measured data), as our goal is to avoid reliance on real measurements and thus train HAR models entirely on synthetic data. Training for IMUGPT and Text2IMU was performed on an entire set of synthetic data, while testing was performed on the entire measured benchmark dataset.

III. Results

Figure 3 shows the results of ten runs for all three benchmark datasets. Across all datasets, models trained with measured data achieved an average accuracy of 76.9% for Activity Set 3, 84.9% for Activity Set 2, and 91.2% for Activity Set 1 (for Activity Set definition see Sec. II-D). Text2IMU data yielded 43.1%, 63.5%, and 79.2% for the three activity sets, whereas IMUGPT⋆ resulted in 14.2%, 22.4%, and 24.7%, respectively. Models trained with measured data performed best in most experiments. Models trained on uncalibrated IMUGPT⋆ data showed the lowest average performance across all experiments and a larger standard deviation in most activity sets compared to Text2IMU and real data. Overall, Text2IMU doubled the accuracy of IMUGPT⋆ across all activity sets.

For Activity Sets 1 and 2 of PAMAP2 and Activity Set 1 of RealWorld, accuracy of Text2IMU was comparable to measured data. While the Text2IMU performance drop compared to measured data for Activity Set 3 of RealWorld was 25.9%, the performance drop decreased to 10.6% on Set 1. For the PAMAP2 dataset, the decrease was 13.2% on Activity Set 2. For Activity Set 1, accuracy reached 94.3%, thus even a 4.8% increase over the average of 89.8% for measured data.

IV. Discussion and Conclusion

We showed effective HAR model training, exclusively using Text2IMU-generated synthetic data. We avoid data mixing or calibration of synthesised data by measured data from the target dataset. Text2IMU yields realistic activity patterns, which is demonstrated by the enhanced HAR performance compared to IMUGPT⋆ (see Fig. 2 and Fig. 3). We attribute the performance increase of Text2IMU to the extended sources of data variation in our approach: multiple surface-based sensor positions were used for each measured IMU instead of joint locations, diverse body model morphologies were included, and various data augmentation strategies were applied. Our results warrant further analysis of the above-mentioned design aspects in more detail to maximise model accuracy.

For Activity Set 3, i.e., all activities of each benchmark dataset, Text2IMU yielded best results on RealWorld (52.6%), followed by WISDM (49.2%), and PAMAP2 (27.4%) (see Fig. 3. We attribute the performance differences to the different activity classes in each dataset. RealWorld and WISDM include rather basic activities, while PAMAP2 comprises rare and complex activities that T2M-GPT was unable to accurately generate, e.g., *Ironing* and *Vacuum Cleaning*. In contrast, Activity Sets 1 and 2 excluded most inaccurate or erroneously generated activities. Thus, HAR accuracy of Text2IMU increased for Sets 1 and 2 substantially. In particular, Activity Set 1 trained with Text2IMU synthetic data showed similar accuracy compared to training on measured data (80.2% vs. 89.7% for Realworld (Fig. 3A), and 94.3% vs. 89.8% for PAMAP2 (Fig. 3B). The similar performances indicate the potential of synthetic data for HAR model training.

The substantial performance increase from Activity Sets 3 to 1 for Text2IMU can be attributed to a current limitation

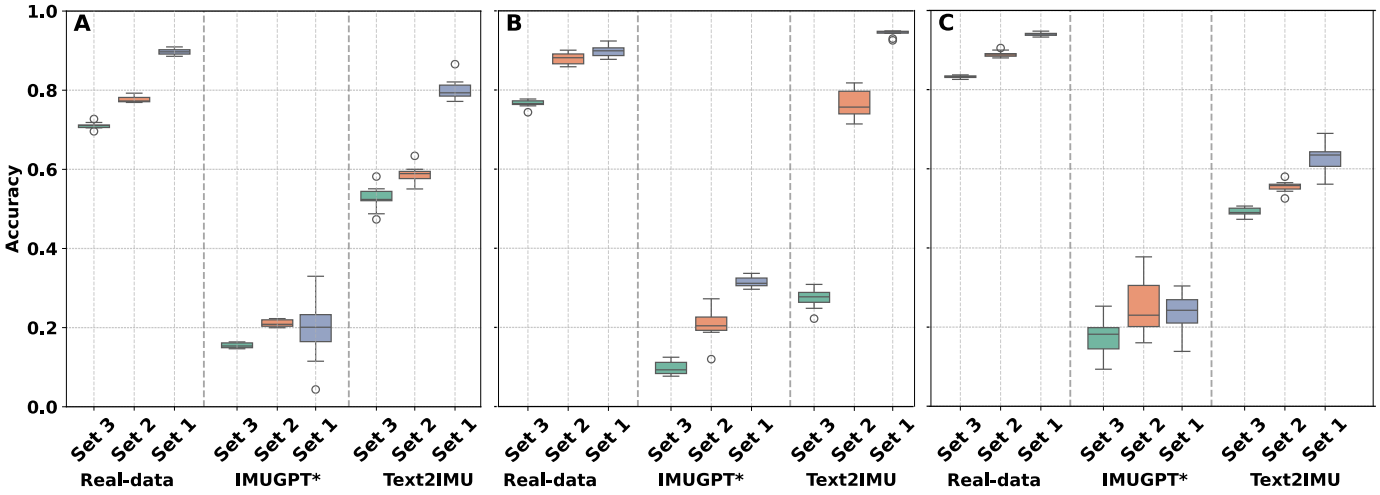

Fig. 3. Accuracy of DeepConvLSTM models trained with measured data, synthetic data obtained by IMUGPT⋆ without calibration, and synthetic data obtained with Text2IMU. Datasets: (A) RealWorld, (B) PAMAP2, and (C) WISDM. Three Activity Sets were evaluated based on quality ratings of five human raters for the activities generated by T2M-GPT (see Sec. II-D): Set 3 includes all activities (A: 8, B: 12, C: 6); Set 2 includes 6 activities (C: 4); Set 1 includes 4 activities (C: 3). IMUGPT⋆ was used without synthetic data calibration, thus served as meaningful comparison to our approach.

of text-based IMU synthesis: While T2M-GPT is a state-of-the-art motion synthesis model, some activities cannot be accurately generated and others are confused, e.g. *Standing* with *Sitting* or *Stairs Up* with *Stairs Down*. Consequently, T2M-GPT may produce erroneous motion patterns and the synthetic IMU data would underperform measured data in HAR tasks. However, we believe that future improvement in motion synthesis models or extensions of the underlying HumanML3D [5] dataset will directly translate into larger activity sets that could be meaningfully used with Text2IMU. Furthermore, future work should systematically evaluate the impact of key synthesis parameters on HAR performance, including the number of generated prompts, body morphology, and virtual IMU placement.

In conclusion, our approach generates realistic motion variations by replicating real world variations: multiple body shape models, multiple virtual sensor placements, combined with synthetic data augmentation. Text2IMU has the potential to facilitate large-scale synthetic data generation and may support the development of foundation-scale machine learning models.

References

[1] F. Gu, M.-H. Chung, M. Chignell, S. Valaee, B. Zhou, and X. Liu, "A Survey on Deep Learning for Human Activity Recognition," *ACM Computing Surveys*, vol. 54, no. 8, pp. 177:1–177:34, Oct. 2021.

[2] J. Seiter, O. Amft, M. Rossi, and G. Tröster, "Discovery of activity composites using topic models: An analysis of unsupervised methods," *Pervasive and Mobile Computing*, vol. 15, pp. 215—227, Dec. 2014.

[3] H. Kwon, C. Tong, H. Haresamudram, Y. Gao, G. D. Abowd, N. D. Lane, and T. Plötz, "IMUTube: Automatic Extraction of Virtual on-body Accelerometry from Video for Human Activity Recognition," *Proceedings of ACM IMWUT*, vol. 4, no. 3, pp. 1–29, Sep. 2020.

[4] L. Uhlenberg, L. O. Haeusler, and O. Amft, "SynHAR: Augmenting human activity recognition with synthetic inertial sensor data generated from human surface models," *IEEE Access*, vol. 12, pp. 194 839–194 858, 2024.

[5] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating diverse and natural 3d human motions from text," in *Proceedings of the IEEE/CVF CVPR*, June 2022, pp. 5152–5161.

[6] J. Zhang, Y. Zhang, X. Cun, Y. Zhang, H. Zhao, H. Lu, X. Shen, and S. Ying, "Generating Human Motion from Textual Descriptions with Discrete Representations," in *2023 IEEE/CVF (CVPR)*. IEEE, pp. 14 730–14 740.

[7] Z. Leng, H. Kwon, and T. Ploetz, "Generating Virtual On-body Accelerometer Data from Virtual Textual Descriptions for Human Activity Recognition," in *Proceedings of the 2023 ACM ISWC*, Oct. 2023, pp. 39–43.

[8] Z. Leng, A. Bhattacharjee, H. Rajasekhar, L. Zhang, E. Bruda, H. Kwon, and T. Plötz, "Imugpt 2.0: Language-based cross modality transfer for sensor-based human activity recognition," *Proc. ACM IMWUT*, vol. 8, no. 3, Sep. 2024.

[9] L. S. S. Ray, B. Zhou, S. Suh, L. Krupp, V. F. Rey, and P. Lukowicz, "Text me the data: Generating Ground Pressure Sequence from Textual Descriptions for HAR." IEEE Computer Society, pp. 461–464.

[10] X. Zuo, S. Wang, J. Zheng, W. Yu, M. Gong, R. Yang, and L. Cheng, "Sparsefusion: Dynamic human avatar modeling from sparse rgbd images," *IEEE Transactions on Multimedia*, vol. 23, pp. 1617–1629, 2021.

[11] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: a skinned multi-person linear model," *ACM Transactions on Graphics*, vol. 34, no. 6, pp. 1–16, Nov. 2015.

[12] X. Zhang, D. Teng, R. R. Chowdhury, S. Li, D. Hong, R. K. Gupta, and J. Shang, "Unimts: Unified pre-training for motion time series," in *NeurIPS*, vol. 37. Curran Associates, Inc., 2024, pp. 107 469–107 493.

[13] S. Pujades, B. Mohler, A. Thaler, J. Tesch, N. Mahmood, N. Hesse, H. H. Bülthoff, and M. J. Black, "The virtual caliper: Rapid creation of metrically accurate avatars from 3d measurements," *IEEE TVCG*, vol. 25, no. 5, pp. 1887–1897, 2019.

[14] L. Uhlenberg and O. Amft, "Comparison of Surface Models and Skeletal Models for Inertial Sensor Data Synthesis," in *BSN '22: Proceedings of the 18th IEEE-EMBS BSN*. Ioannina, Greece: IEEE, Sep. 2022.

[15] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 22 140, pp. 55–55, 1932.

[16] T. Sztyler and H. Stuckenschmidt, "On-body localization of wearable devices: An investigation of position-aware activity recognition," in *2016 IEEE PerCom*. Sydney, Australia: IEEE, Mar. 2016, pp. 1–9.

[17] A. Reiss and D. Stricker, "Introducing a New Benchmarked Dataset for Activity Monitoring," in *2012 16th International Symposium on Wearable Computers*. ACM, Jun. 2012, pp. 108–109, iSSN: 2376-8541.

[18] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, Mar. 2011.

[19] F. J. Ordóñez and D. Roggen, "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition," *Sensors*, vol. 16, no. 1, p. 115, Jan. 2016, number: 1 Publisher: Multidisciplinary Digital Publishing Institute.