Enhancing Point-to-Point Upper-Limb Movement Detection Using Inertial Images and Deep Learning

Meghana Sai Veligatla
Manning CICS
University of Massachusetts Amherst
Amherst, USA
mveligatla@umass.edu

Thashmitha Bangalore Shekar
Manning CICS
University of Massachusetts Amherst
Amherst, USA
tbs@umass.edu

***Abstract*—Human Activity Recognition (HAR) using wearable inertial sensors is a critical yet challenging task, particularly in rehabilitation, where labeled data are limited and time-series signals are complex to model. This study describes an inventive pipeline for transforming multi-axis inertial sensor data (X, Y, and Z) into time-invariant RGB inertial images using plane-wise projections (XY, XZ, and YZ), allowing the usage of pre-trained deep learning models such as EfficientNet-B0, which was originally trained on ImageNet.We validate the technique on two datasets: (1) detecting point-to-point (P2P) upper-limb motions in stroke survivors, which is critical for motor recovery assessment, and (2) classifying everyday activities (walking, running, sitting, standing, and climbing stairs) in the RealWorld HAR database. Using stratified subject splits and Leave-One-Subject-Out Cross-Validation (LOSOCV), the system obtains accuracy of 84-87% on the stroke dataset and 85-90% on the RealWorld HAR dataset.This study shows that translating inertial signals into visual representations enables robust, transferable feature extraction and generalization in clinical and real-world scenarios, providing an efficient framework for future HAR applications.**

***Index Terms*—Human Activity Recognition, Inertial Images, Deep Learning, EfficientNet, Stroke Rehabilitation, Transfer Learning, Cross-Validation, Preprocessing, Model Evaluation**

I. Introduction

Human Activity Recognition (HAR) utilizing wearable inertial measurement units (IMUs) is an emerging technology for improving healthcare monitoring, rehabilitation evaluation, fitness tracking, and human-computer interaction systems [3], [7]. HAR systems capture fine-grained motion signals from accelerometers, gyroscopes, and magnetometers, allowing for a deep understanding of physical activities across various populations. Accurately identifying point-to-point (P2P) upper-limb motions is crucial in clinical contexts, such as stroke rehabilitation, to monitor motor recovery, evaluate the success of therapeutic interventions, and guide adaptive treatments [5].

HAR experiences several challenges, including limited labeled datasets, particularly in clinical populations, significant variability in movement patterns across individuals, and the inherent complexity of high-dimensional, time-dependent sensor signals [4], [6]. Traditional HAR techniques rely on hand-crafted features and shallow models, which typically fail to generalize well, especially when applied to populations with damaged or abnormal motor functions, such as stroke survivors [7]. Deep learning techniques, specifically convolutional neural networks (CNNs), have revolutionized HAR by enabling automatic feature extraction. However, training such models from scratch is challenging for compact, domain-specific datasets [4].

This work proposes a new image-based HAR pipeline that converts multi-axis IMU time-series data into RGB inertial images, allowing for the application of pre-trained CNNs like EfficientNet-B0 [1]. By expressing temporal sensor input as static 2D visual patterns using plane-wise projections (XY, XZ, YZ), the approach takes advantage of the great generalization capabilities of models trained on large-scale visual datasets such as ImageNet. Our pipeline is tested on two datasets: a clinical stroke dataset for detecting P2P upper-limb movements, and the RealWorld HAR dataset for classifying five common activities [2]. To establish a rigorous and realistic evaluation, we use stratified subject splits and Leave-One-Subject-Out Cross-Validation (LOSOCV), which reduces data leakage and assesses generalizability to unseen individuals.

This work makes three significant contributions: (1) we introduce an intriguing transformation technique that encodes IMU signals as RGB inertial images, (2) we illustrate the effective application of pre-trained EfficientNet-B0 for sensor-based HAR tasks with minimal fine-tuning, and (3) we conduct a thorough evaluation on both clinical and real-world datasets, achieving high accuracies (84-87% on stroke, 85-90% on RealWorld HAR) and demonstrating the approach's scalability. This study adds to the increasing literature on transfer learning and cross-domain adaptation for wearable sensor applications, paving the way for future research in healthcare, sports science, and human-centered computing [3], [5], [6].

II. Related Work

Wearable sensor-based Human Activity Recognition (HAR) research has progressed from traditional statistical techniques to advanced deep learning algorithms [3], [7]. Early HAR systems used hand-crafted feature extraction to calculate descriptive statistics (mean, variance) and frequency-domain characteristics, which were then fed into classical classifiers like Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), or Random Forests [7]. Although these approaches worked well in controlled or homogenous datasets, they frequently struggled to generalize across subjects due to inter-individual variability, needing extensive domain expertise to tune features and thresholds successfully [4], [7].

Deep learning models, such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and 1D convolutional neural networks (CNNs), can now process raw time-series data directly [4], [5]. However, these approaches come with significant limitations: Large, well-annotated datasets are typically needed in clinical settings, especially in stroke rehabilitation where gathering high-quality, labeled movement data is labor-intensive and subject to ethical constraints [6]. Deep time-series models can be computationally demanding and prone to overfitting when applied to small or imbalanced datasets [4], [5].

To address these issues, recent studies have explored translating IMU time-series signals into 2D image-like representations, including spectrograms, recurrence plots, Gramian Angular Fields, and Markov Transition Fields [5], [6]. These changes enable researchers to leverage the strength of pre-trained 2D convolutional models, such as those trained on ImageNet. This reduces the dependency on huge labeled datasets and leverages robust visual feature extractors [1], [4]. However, many of these picture transformations (e.g., spectrograms) may sacrifice critical temporal or spatial correlations, whereas others (e.g., recurrence plots) are computationally expensive to construct and scale poorly to real-time applications.

Our approach extends the idea of inertial images presented in [6], generating plane-wise projections (XY, XZ, YZ) of multi-axis IMU data to construct color-encoded (RGB) images. This approach preserves both the inter-axis relationships and the temporal continuity of the data in a time-invariant visual format, making it ideal for future 2D CNN designs. Our pipeline enhances classification accuracy and reduces computing costs by utilizing EfficientNet-B0 [1], a scalable and efficient vision model pre-trained on ImageNet. Our focus on stroke rehabilitation addresses an underexplored area of HAR, where altered movement patterns pose specific obstacles that general-purpose HAR research normally do not address [5], [6].

Furthermore, unlike previous research, which mostly evaluates models using random or stratified splits, we rigorously use Leave-One-Subject-Out Cross-Validation (LOSOCV), which better replicates real-world deployment by examining the system's capacity to generalize to completely unknown individuals. Our contribution is particularly significant for clinical applications where patient-specific calibration is typically problematic [4], [6].

III. Methodology

A. Datasets

To evaluate our approach, we rely on two datasets that include both clinical and everyday contexts, ensuring the pipeline's adaptability and transferability.

1) Stroke Dataset: This dataset includes wrist-mounted IMU recordings from 18 stroke survivors taken at 100 Hz. The data are classified into five activity classes: point-to-point (P2P, 1789 samples), wrist rotation (WR, 153 samples), repetitive (Rep, 245 samples), hand-to-mouth (H2M, 85 samples), and discard (696 samples), for a total of 2968 segments. This dataset exhibits common issues in clinical HAR tasks, such as high inter-subject variability, minor motor deficits, and class imbalance where P2P activities dominate due to their central role in rehabilitation exercises.

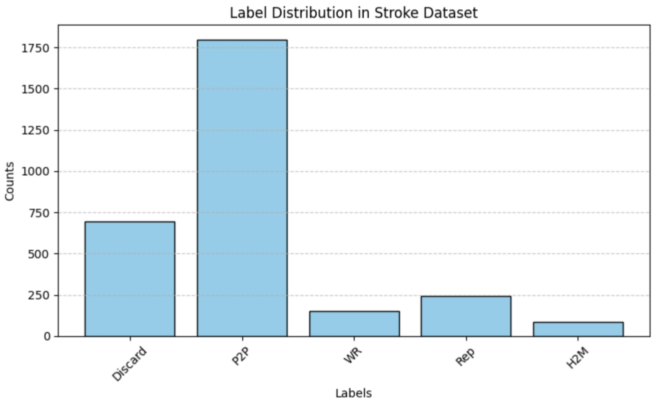

Fig. 1. Stroke Data Distribution

2) RealWorld HAR Dataset: The RealWorld HAR dataset [2] includes almost 20,000 samples from 7 healthy participants performing five daily activities: walking, running, sitting, standing, and climbing stairs. This dataset serves as a balanced, high-volume benchmark for determining the pipeline's generalizability on everyday activities

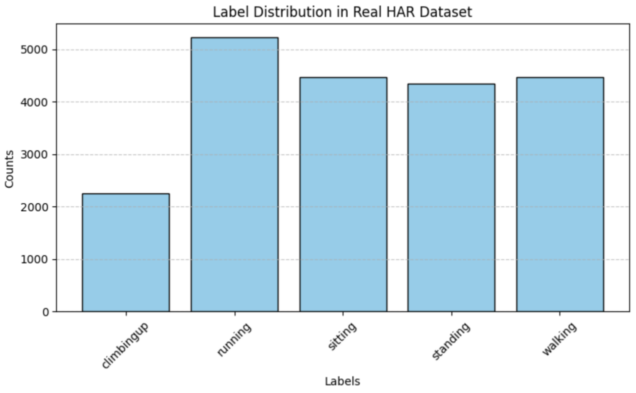

Fig. 2. RealWorld HAR Data Distribution

B. Data Preprocessing

Given the complexities of raw IMU data, we employ a multi-step preparation pipeline before feeding it into the model. Raw IMU signals from the X, Y, and Z axes are preprocessed to reduce noise, normalize the data, and get ready for inertial image production. The Preprocessing pipeline consists of the following major steps:

Filtering: To smooth raw IMU data, we use a 4th-order Butterworth low-pass filter with a cutoff frequency of 5 Hz. This filtering step is necessary because wearable sensor data frequently contains high-frequency noise caused by sensor jitter, surroundings vibrations, or unexpected non-representative spikes. By deleting these extraneous high-frequency components, we preserve the genuine underlying movement patterns while ensuring that the resulting inertial images reflect meaningful motion dynamics rather than noise artifacts. This increases both the visual quality of the inertial pictures and the downstream model's capacity to learn strong, discriminative features.

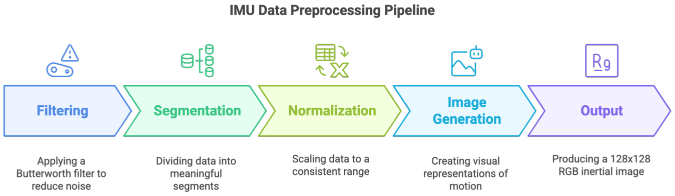
IMU Data Preprocessing Pipeline
Filtering
Segmentation
Normalization
Image Generation
Output
Applying a Butterworth filter to reduce noise
Dividing data into meaningful segments
Scaling data to a consistent range
Creating visual representations of motion
Producing a 128x128 RGB inertial image

Fig. 3. IMU Data Preprocessing Pipeline

Segmentation: We used multiple segmentation strategies that are specific to each dataset's properties.

- *For the stroke dataset*, we utilize fully annotated segmentation, extracting segments entirely on expert-labeled start and finish times for each activity. This method ensures that each segment accurately corresponds with clinically important motions while reducing contamination from irrelevant or transitional motion. By relying on these tidy, restricted activity windows, we feed the model high-quality, task-specific samples, which is especially crucial for detecting subtle or impaired motions in stroke survivors.
- *For the RealWorld HAR dataset*, which lacks detailed time-aligned annotations, we use a sliding window technique with a 2-second window and 50% overlap. This approach splits the continuous sensor stream into overlapping segments, increasing the amount of training examples and collecting a variety of temporal circumstances. The overlap allows the model to learn robust, invariant properties by exposing it to slightly altered copies of the same activity, improving its generalization to changes in how healthy persons complete ordinary tasks.

Normalization: In order to align with the pixel intensity scale utilized in the 128x128 inertial pictures, the X, Y, and Z axis signals for both datasets are individually normalized to the range [0,127]. This step guarantees that sensor range, subject size, and motion amplitude do not dominate the data, resulting in a consistent representation across all subjects and activities. We standardize the data input by normalizing the raw sensor readings, resulting in comparable images and guaranteeing that the deep learning model focuses on significant movement patterns rather than absolute signal magnitudes.

Image Generation: Movement trajectories are plotted as lines on a white canvas:

- *Red channel*: depicts the XY plane, highlighting movements that combine lateral and vertical changes.
- *Green channel*: maps the XZ plane, emphasizing the interaction between depth and lateral motions.
- *Blue channel*: reflects the YZ plane, demonstrating the interaction of vertical and depth movements.

This multiplane encoding preserves cross-axis linkages and trajectory patterns in a time-invariant visual representation, making movement data suitable for 2D convolutional neural networks. By expressing temporal sensor data as static visual patterns, we enable the use of sophisticated pre-trained picture models (such as EfficientNet-B0), allowing the system to benefit from extensive spatial feature extraction capabilities without requiring enormous labeled sensor datasets.

Output: The final outcome is a 128x128 RGB inertial image that preserves both spatial relationships and temporal motion patterns, resulting in a compact, time-invariant visual representation of the recorded activity.

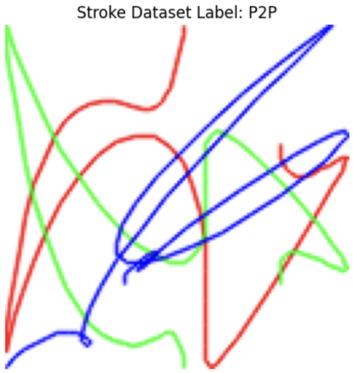

Fig. 4. Sample Stroke Data RGB Image

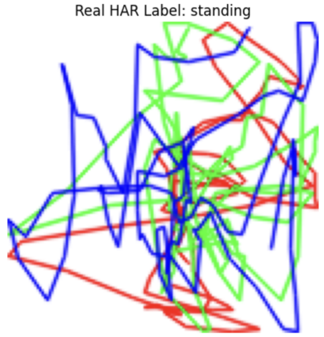

Fig. 5. Sample HAR Data RGB Image

C. Data Augmentation

In the current investigation, we purposefully avoided using data augmentation techniques in order to assess how well the suggested pipeline operates on its own, without introducing false variance. To better understand the model's baseline learning behavior, we excluded geometric transformations (such as random rotations, flips, and scaling), photometric adjustments (brightness and contrast changes), and synthetic oversampling methods (such as SMOTE for minority classes). Future research will look at merging these augmentation methodologies to improve resilience and performance in diverse and imbalanced data sets.

D. Model Architecture

We employ EfficientNet-B0, a lightweight convolutional neural network with around 5.3 million parameters, and use its pre-trained ImageNet weights for robust, general-purpose picture feature extraction [1]. This model is ideal for our inertial picture pipeline because of its compound scaling technique, which effectively balances the depth, width, and resolution of the network.

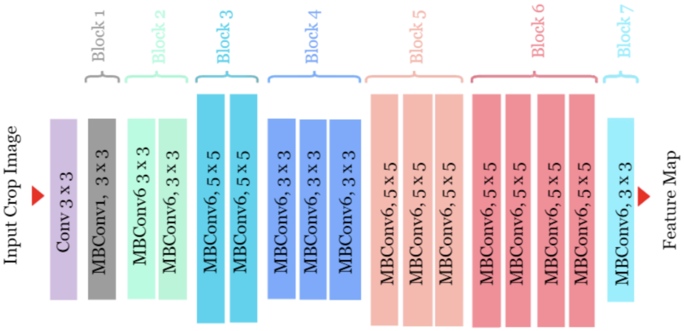

Fig. 6. Model Architecture

Our implementation includes the following critical components:

- **Backbone**: We maintain the pre-trained EfficientNet-B0 feature extractor (`model.features`) as the backbone. This component provides rich, transferable representations that capture visual patterns such as edges, textures, and spatial arrangements, allowing the model to interpret the time-invariant patterns encoded in inertial images.
- **Adaptive Pooling**: we use an adaptive average pooling layer (`nn.AdaptiveAvgPool2d(1)`) to compress the spatial dimensions of the feature maps into a global descriptor of fixed size. This step ensures that the next classifier receives a compact, consistent input regardless of image size.
- **Classifier Head**: We replace EfficientNet's original classification layer with a custom fully connected layer (`nn.Linear`) that matches the number of activity classes (five in both the stroke and real-world HAR datasets). In our system, only the classifier head is fine-tuned during training, considerably reducing the number of trainable parameters and limiting the danger of overfitting, which is especially crucial considering the stroke dataset's small size.
- **Regularization**: To further prevent overfitting, we apply a 0.5 dropout rate and L2 weight regularization (weight decay = 0.01) to the classifier head. These strategies encourage generalization by minimizing the possibility that the model would memorize the training data rather than discover significant patterns.
- **Input handling**: To satisfy the input size expectations of EfficientNet-B0, the generated 128x128 inertial pictures are resized to 224x224 pixels by bilinear interpolation. The images are then transformed to PyTorch tensors and reshaped into the (batch, channels, height, width) structure, which is the typical convention for convolutional models.
- **Training Pipeline**: Our `EfficientNetClassifier` class encompasses the backbone, pooling, and classification components, and integrates easily into the training loop. We optimize the model using the Adam optimizer (starting learning rate = 0.001) in conjunction with a cosine annealing scheduler to gradually reduce the learning rate over time while encouraging steady convergence. To efficiently manage GPU memory during cross-validation, we employ Python's garbage collector to clean the CUDA cache between runs, resulting in reliable execution on Colab.

Overall, this architecture is intended to strike a compromise between efficiency, adaptability, and generalization, making it ideal for learning from time-invariant visual representations of multi-axis inertial sensors.

E. Training and Evaluation

We evaluate model performance using two complimentary evaluation methodologies meant to balance practicality and robustness:

- **Stratified subject splits**: When multiple subjects are available, stratified subject splitting offers a practical way to divide the data into training (80%) and testing (20%) sets while maintaining balanced class distributions across folds. This strategy assures that each split includes representative samples from all activity classes, which reduces the possibility of class imbalance skewing performance findings. We use this function to randomly choose subjects over multiple folds while keeping subject identity and label proportions. Stratified splits give a quick, balanced estimation of the model's generalization ability across a diversified subject pool.
- **Leave-One-Subject-Out Cross-Validation (LOSOCV)** For a more rigorous and realistic evaluation, we limit the test set to one subject at a time while training on the others. This technique directly mirrors real-world deployment, in which the system must generalize to previously unknown users, making it particularly useful in personalized or clinical applications.

Training lasts 10 epochs and employs the Adam optimizer with an initial learning rate of 0.001, guided by a cosine annealing scheduler that gradually reduces the learning rate over time to promote stable convergence. We track and plot training and validation accuracy and loss for each epoch, allowing us to observe learning progress and spot potential underfitting or overfitting.

Performance evaluation is centered on confusion matrices and multiclass ROC curves, which provide extensive class-wise insights into the model's discriminative capacity and mistake patterns. We prioritize these visual measures to identify where the model succeeds or fails, rather than reporting aggregate metrics such as precision, recall, or F1-score.

IV. Results

A. Results on Stroke Dataset

Our proposed pipeline achieves an average accuracy of 85% across stratified subject splits and Leave-One-Subject-Out Cross-Validation (LOSOCV), exhibiting robust generalization despite the limitations provided by the stroke dataset's structure.

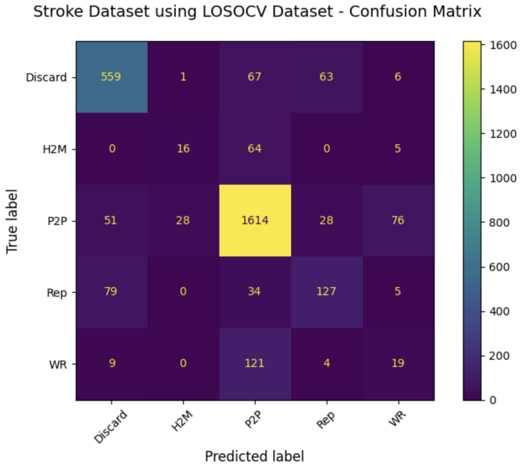

Fig. 7. Stroke Data LOSOCV Confusion matrix

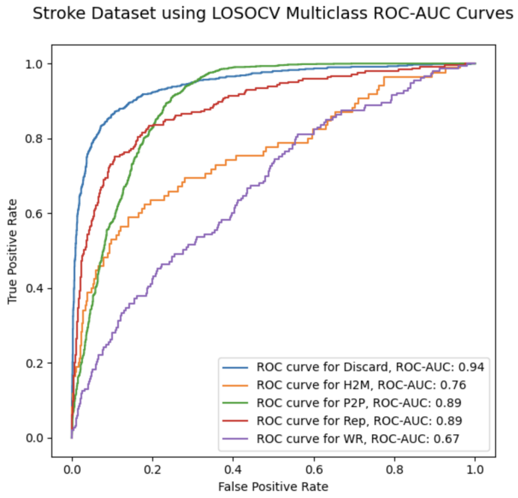

Fig. 8. Stroke Data LOSOCV ROC-AUC Curves

As seen in Figure 7, the LOSOCV confusion matrix demonstrates good recognition of the dominant P2P class (1789 samples), but also reveals confusion between P2P and minority classes, such as WR and H2M. This is obvious given the inherent class imbalance and the fact that many inertial pictures across these classes have relatively similar motion patterns, particularly for restricted or brittle stroke motions. Despite the hurdles, the multiclass ROC-AUC curves in Figure 8 demonstrate that the model maintains high discriminative capacity, reaching ROC-AUC scores of 0.89 for P2P and 0.94 for the Discard class, although reduced separability is observed for WR (0.67) and H2M (0.76).

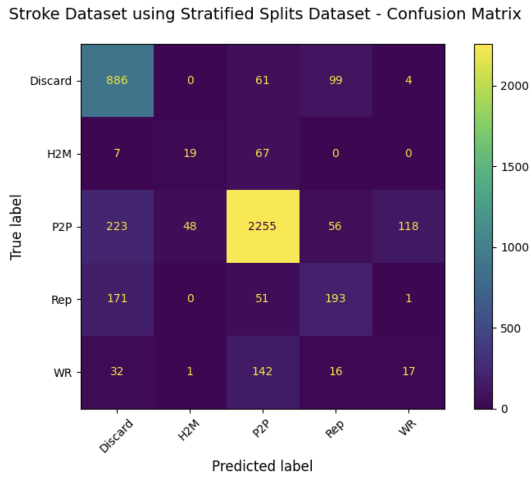

Fig. 9. Stroke Data Stratified Confusion matrix

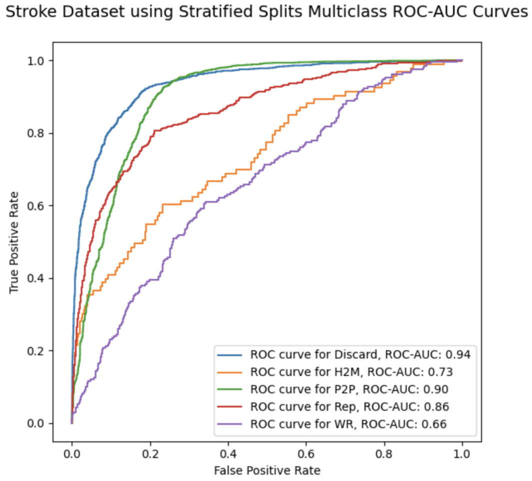

Fig. 10. Stroke Data Stratified ROC-AUC Curves

In Figure 9, the stratified splits confusion matrix benefits from more balanced subject representation per fold, which helps reduce subject-specific bias and leads to slightly better ROC-AUC results overall (Figure 10), with P2P reaching 0.90 and Rep achieving 0.86. This method guarantees that minority classes are constantly included in both training and testing, providing the model with additional exposure across classes. However, minority classes such as H2M and WR continue to have lower ROC-AUC scores (0.73 and 0.66) and higher misclassification rates, highlighting the persistent challenge of class imbalance and the difficulty of distinguishing visually similar inertial patterns, particularly for tiny stroke motions. These data indicate that, while stratified splitting improves overall performance, other measures such as augmentation or class balancing may be required to improve outcomes for underrepresented classes.

Importantly, these findings demonstrate that, even without data augmentation, the pipeline effectively learns discriminative features from inertial images using pre-trained visual models such as EfficientNet-B0. Despite constraints such as class imbalance and small movement fluctuations, the system maintains steady and constant high performance across both evaluation methodologies, highlighting its potential for dependable activity recognition in stroke therapy. The code also includes documentation of the training loss and accuracy progression over epochs, which can be used to conduct additional learning dynamics research.

TABLE I
ROC-AUC SCORES FOR STROKE DATASET (LOSOCV VS. STRATIFIED SPLITS)

Class	LOSOCV	Stratified Splits
Discard	0.94	0.94
H2M	0.76	0.73
P2P	0.89	0.90
Rep	0.89	0.86
WR	0.67	0.66

While both LOSOCV and stratified splits produce fairly comparable average ROC-AUC scores across most classes, the evaluation procedures are fundamentally different. Stratified splits are effective when there are many subjects because they ensure balanced class distributions throughout training and test sets. However, stratified splitting might sometimes overstate generalization since the model may encounter data from the same subject in both the training and test sets, resulting in subject-level leakage. LOSOCV, on the other hand, is stricter: it excludes entire subjects, ensuring that the model is evaluated only on previously unknown individuals. This makes LOSOCV a more rigorous and realistic evaluation tool for determining how effectively the model generalizes to new users, particularly in sensitive clinical settings such as stroke rehabilitation.

B. Results on RealWorld HAR Dataset

Our proposed pipeline delivers an average accuracy of 88%-90% across both stratified subject splits and Leave-One-Subject-Out Cross-Validation (LOSOCV), exhibiting strong generalization despite the limitations provided by the RealWorld HAR dataset's structure.

The LOSOCV confusion matrix, depicted in Figure 11, accurately classifies main activities such as running and sitting using strong diagonal counts. However, significant confusion exists between similar actions such as climbing up and walking or standing and sitting, most likely due to overlapping motion characteristics in the inertial data. The ROC-AUC curves in Figure 12 show a considerable class separation, with sitting having the greatest ROC-AUC of 0.91, followed by walking at 0.88 and running at 0.86. Climbing (0.85) and standing (0.81) have somewhat lower ROC-AUC values, showing that they remain difficult to differentiate.

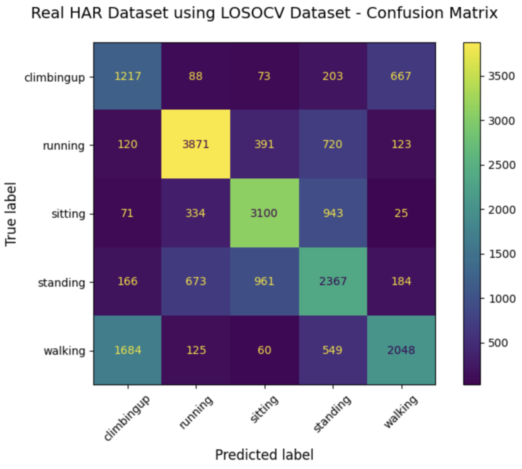

Fig. 11. RealWorld HAR Data LOSOCV Confusion matrix

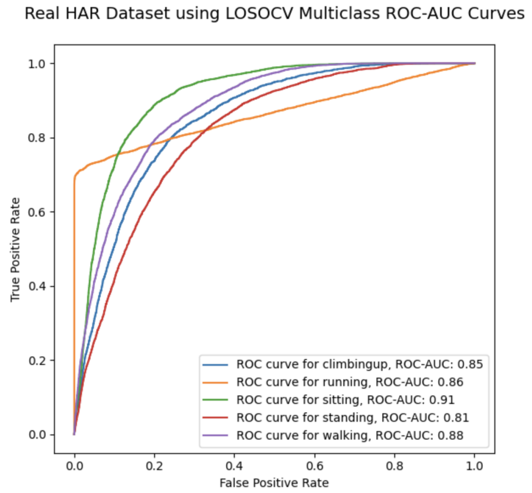

Fig. 12. RealWorld HAR Data LOSOCV ROC-AUC Curves

The stratified split evaluation (Figures 13 and 14) shows similar trends, with sitting attaining a top ROC-AUC of 0.91 and walking achieving 0.90. Although stratified splits give balanced folding, some dynamic or transitional tasks, such as standing (0.77) and running (0.84), remain fairly difficult due to motion overlap.

The results obtained show that the pipeline generalizes effectively for ordinary human activity recognition tasks, with especially excellent performance on static and repetitive activities, while leaving room for improvement on transitional movements.

On the RealWorld HAR dataset, the pipeline achieves good

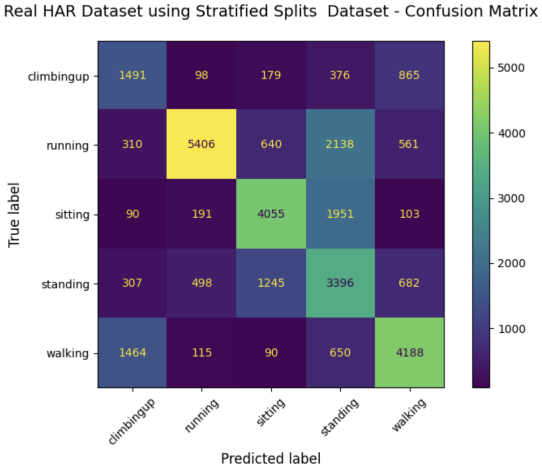

Fig. 13. RealWorld HAR Data Stratified Confusion matrix

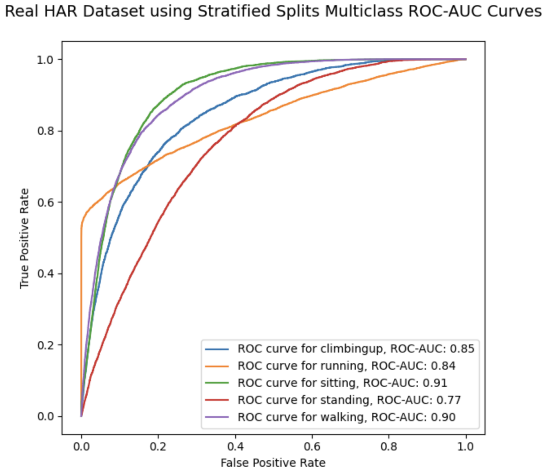

Fig. 14. RealWorld HAR Data Stratified ROC-AUC Curves

and consistent ROC-AUC performance under both LOSOCV and stratified subject splits, with average values ranging from 0.77 to 0.91 across classes. Notably, sitting and walking regularly earn the greatest ROC-AUC scores (above 0.9), indicating clear separability in their inertial signatures, whereas standing has the lowest scores (0.77-0.81), most likely due to postural similarities to other static activities. Overall, all assessment procedures produce comparable results, but stratified splits slightly improve the walking class, demonstrating that bigger subject mixtures can occasionally increase class balancing effects while preserving generalization.

TABLE II
ROC-AUC Scores for RealWorld HAR Dataset (LOSOCV vs. Stratified Splits)

Class	LOSOCV	Stratified Splits
ClimbingUp	0.85	0.85
Running	0.86	0.84
Sitting	0.91	0.91
Standing	0.81	0.77
Walking	0.88	0.90

C. *Training Dynamics*

Over ten epochs, the training dynamics of both the stroke dataset and the RealWorld HAR dataset demonstrate smooth and stable convergence. In all runs, the model continuously shows a clear drop in training loss and a steady improvement in accuracy, indicating good learning and dependable optimization behavior.

For the stroke dataset, the final training accuracy is typically 93-95%, with loss decreasing progressively over epochs, demonstrating that the EfficientNet-B0 backbone effectively captures meaningful features from inertial picture inputs. Similarly, for the RealWorld HAR dataset, training accuracy improves steadily, reaching 90-92% by the end of training. Notably, these results were obtained without the use of data augmentation or complex regularization, demonstrating the effectiveness of the combined image-based representation and pre-trained model backbone.

The training patterns in both datasets show no significant signs of overfitting, indicating that the use of frozen feature extractors, combined with a lightweight fine-tuned classification head and modest regularization (such as dropout and weight decay), achieves a strong balance between learning capacity and generalization. These tendencies, as reported in the training logs and illustrated in the epoch-wise charts, highlight the proposed pipeline's robustness and scalability across clinical and daily activity detection tasks.

V. Discussion

The suggested pipeline performs well on both clinical and real-world datasets, demonstrating the efficacy of translating multi-axis inertial data into image representations and including transfer learning. The stroke dataset's average accuracy of 85% is chiefly limited by class imbalance and visual resemblance within classes such as P2P, WR, and H2M, making fine-grained separation difficult. The RealWorld HAR dataset, on the other hand, yields greater accuracy (88-90%), as well as stronger ROC-AUC scores, thanks to its more balanced class distribution and unique movement patterns between activities such as walking, sitting, and climbing. This increased class diversity and clearer activity separation enable the pipeline to achieve improved class-specific performance, demonstrating its flexibility to a wide range of application domains.

Using pre-trained EfficientNet-B0 considerably decreases the need for large labeled datasets and shortens training time, making this approach ideal for clinical or resource-constrained settings. Importantly, the implementation of Leave-One-Subject-Out Cross-Validation (LOSOCV) assures that the system generalizes effectively to new users, which is essential for real-world deployment in tailored rehabilitation, fitness tracking, and aged care systems. Furthermore, the plane-wise projection method preserves crucial movement properties, enabling the model to detect meaningful spatial patterns hidden in inertial signals.

A. Practical implications

In clinical practice, this pipeline could aid in the automated assessment of point-to-point (P2P) upper-limb movements, thereby offering objective, consistent, and scalable support for stroke rehabilitation programs. Reducing reliance on manual therapist observation has the potential to free up clinical resources and improve patient monitoring outside of clinical settings. The system's outstanding performance on the Real-World HAR dataset indicates that it can be used to applications like as personal fitness tracking, daily activity monitoring, and geriatric care, where continuous, non-intrusive sensing via wearable IMUs can enable early intervention and promote better lives.

B. Challenges

Despite its strengths, the pipeline faces several challenges. The stroke dataset's class imbalance results in inferior performance for minority classes such as H2M and WR, as seen in Table I. Visual similarity across inertial pictures complicates misclassifications, especially for hindered or faint stroke motions. Furthermore, creating inertial pictures and fine-tuning EfficientNet-B0 demand processing resources, which may limit its use on low-power devices. Another problem is susceptibility to sensor placement and orientation; while we assume consistent IMU installation, deviations in real-world settings may add noise and impact generalizability. Furthermore, while LOSOCV enhances user-level generalization, variations in user movement styles, device kinds, and environmental variables may present adaptation issues when moving beyond the training data domain.

VI. Limitations

The pipeline has numerous constraints that deserve consideration:

- **Class Imbalance**: The stroke dataset's uneven distribution, dominated by P2P movements, pushes the model towards majority classes and affects performance on underrepresented categories like WR and H2M.
- **Visual Similarity**: Inertial images from similar movement patterns (e.g., repetitive hand movements vs. point-to-point motions) may appear visually identical, reducing the classifier's fine-grained classification capabilities.
- **Computational Complexity**: The image creation and EfficientNet-B0 inference need significant processing resources, making real-time or mobile deployment hard without additional optimization or model compression.
- **Generalizability**: Although LOSOCV ensures subject-level generalization, the pipeline has not yet been validated on completely different clinical cohorts (e.g., patients from different rehabilitation centers, varying age groups, or cultural backgrounds), which may impact transferability.
- **Temporal Information Loss**: The current approach focuses mostly on spatial information acquired within inertial images, thereby disregarding temporal or sequential patterns that could contain significant hints, especially for complicated or composite activities.
- **The Dependence on Annotation Quality**:The pipeline assumes that the annotated segments are accurate and representative. Any flaws or inconsistencies in the annotation process (for example, mislabeled or unclear activity windows) may propagate downstream, affecting model learning and evaluation.
- **Sensitivity to Sensor Variability**: Differences in IMU sensor types, calibration, or location across datasets or use cases may result in variability that the current model is not explicitly designed to address.

VII. Future Work

Future modifications could solve present constraints and broaden the pipeline's usefulness by:

- **Advanced Data Augmentation**: Using generative adversarial networks (GANs) or other synthetic data generation approaches to generate balanced and diverse inertial picture samples may increase minority class representation, especially for underrepresented stroke activities such as H2M and WR. This would help to reduce class imbalance and increase generalization.
- **Temporal Modeling**: Using architectures that explicitly capture temporal dynamics, such as combining convolutional neural networks (CNNs) with recurrent neural networks (RNNs), temporal convolutional networks (TCNs), or Transformer-based models, may improve the recognition of complex or subtle motion sequences that spatial-only representations may miss.
- **Edge Deployment Optimization**: Using model compression techniques like quantization, pruning, and knowledge distillation can reduce model size and computational demand, allowing for real-time deployment on resource-constrained edge devices such as wearables or smartphones for continuous monitoring.
- **Multi-Sensor Fusion**: Integrating signals from additional body-mounted IMUs (e.g., elbow, shoulder, or torso) or combining inertial data with complementary modalities (e.g., electromyography or vision) can provide richer, multimodal insights into movement patterns, improving classification accuracy and robustness.
- **External and Cross-Population Validation**: Testing the pipeline on external datasets from diverse clinical cohorts, age groups, and environmental settings would be critical to validate its generalizability and ensure that the system performs reliably across different real-world populations and use cases.

  User Personalization: Adaptive learning techniques can be used to personalize models for individual users, boosting performance in real-world applications such as rehabilitation monitoring or fitness tracking.
- **Explainability and Interpretability**: Developing methods to visualize or explain the model's decision-making process (e.g., using attention mechanisms or saliency maps) could increase trust and transparency, especially in

sensitive clinical applications where understanding why a system made a certain classification is critical.

VIII. Conclusion

This paper describes a new human activity recognition (HAR) pipeline that converts IMU time-series data into RGB inertial pictures, allowing the use of pre-trained EfficientNet-B0 for efficient and transferable feature extraction. The pipeline performs well, with 85% accuracy on the clinical stroke dataset and 88-90% on the RealWorld HAR dataset, exhibiting substantial generalization across both healthcare and daily activities domains. By utilizing transfer learning, the system tackles the issue of labeled data scarcity, making it a promising alternative for scalable HAR applications.

Aside from its technical accomplishments, this study demonstrates the broader possibility of merging wearable sensors with deep learning to support rehabilitation, fitness tracking, and senior care. The use of LOSOCV allows generalization to unseen users, which is critical for real-world implementation, while the system's versatility encourages future development into adjacent areas such as sports surveillance or workplace safety.

Future work will address class imbalance, incorporate temporal modeling to improve sequence understanding, and optimize the pipeline for low-power, real-time application. Expanding to multi-sensor fusion and verifying across varied demographics would strengthen the system's robustness and practical utility.Overall, this study establishes the groundwork for future HAR systems that combine wearable sensing, visual deep learning, and scalable deployment to improve health and activity monitoring in real-world settings.

References

[1] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[2] T. Sztyler and H. Stuckenschmidt, "On-body localization of wearable devices: An investigation of position-aware activity recognition," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2016, pp. 1–9.

[3] A. Wang, G. Chen, and C. Yang, "A survey on human activity recognition using wearable sensors," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1323–1350, 2021.

[4] J. Wang, Y. Chen, and S. Hao, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, 2019.

[5] X. Li, Y. Zhang, and I. Marsic, "Deep learning for human activity recognition with wearable sensors," in *Proc. IEEE Int. Conf. Healthcare Informat.*, 2018, pp. 1–6.

[6] H. F. Nweke, Y. W. Teh, and G. Mujtaba, "Data fusion and multiple classifier systems for human activity detection and classification," *Comput. Sci. Rev.*, vol. 34, pp. 100–114, 2019.

[7] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 1192–1209, 2013.