Negative Results in IMU-based Skiing Style Recognition: Skill Level Distribution Shift

Behrooz Azadi
Pro2Future GmbH
Altenberger Strasse 69, 4040 Linz, Austria
behrooz.azadi@pro2future.at

Michael Haslgrübler
Pro2Future GmbH
Altenberger Strasse 69, 4040 Linz, Austria
michael.haslgruebler@pro2future.at

Alois Ferscha
Institute of Pervasive Computing
Johannes Kepler University
Altenberger Straße 69, 4040 Linz, Austria
ferscha@pervasive.jku.at

Abstract

Human Activity Recognition (HAR) models often struggle with generalization, meaning they perform poorly when deployed in new, unseen environments or with different users than those in the training set. This poor generalization occurs because HAR models often experience variability in sensor data due to differences in individuals (e.g., age, gender, physical characteristics) and environmental factors (e.g., sensor placement, lighting conditions, background noise), which cause distribution shift. All mentioned forms of variability can appear in recreational alpine skiing. However, skiers' skill is a factor that has been less studied and highlighted by scholars. In this study, we explain why a model struggles to generalize across a mixed-skill dataset. We employed an autoencoder-based multi-task learning model, which, despite achieving state-of-the-art on standard HAR datasets and promising results on a skiing dataset, failed to generalize in a real-world alpine skiing setting. We identified and quantified a skill-related distribution shift as the cause; low-skilled and experienced skiers occupy distinct regions in latent feature space with Wasserstein-1 distances increasing from 5.39 to 41.85 for the most basic to the most advanced skiing technique.

1 Introduction

Human Activity Recognition (HAR) is a field of research within pervasive computing and human-computer interaction with applications in domains such as healthcare Serpush et al. (2022), sports Hoelzemann et al. (2023), and manufacturing Sopidis et al. (2023). To identify, categorize, and evaluate human activities, HAR relies on machine learning to analyze data captured by visual and wearable sensors.

Although HAR models trained on publicly available datasets achieved the state-of-the-art, activity recognition in real-world scenarios can be demanding due to the variability in the dataset and uncontrolled factors Gil-Martín et al. (2023); Jimale and Mohd Noor (2023). In many real-world scenarios, the test and training sets do not come from the same distribution Koh et al. (2021), which causes a significant performance decline when facing unseen data. This phenomenon, known as distribution shift, causes models to fail when deployed outside their training distribution. An example of such a case is when users perform an activity differently Kreil et al. (2016). For instance, Jimale and

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Learning to Sense.

Mohd Noor (2023) reported a noticeable recognition drop for ML and DL models due to age-related shifts in the dataset.

The other variability in wearable sensor-based HAR is due to sensor placement and orientation, referred to as wearing variability Min et al. (2019), which can cause a distribution shift even for the same subject across distinct recording sessions. Gil-Martín et al. Gil-Martín et al. (2023) and Khaked et al. Khaked et al. (2023) examined the impact of orientation change using transformation and real-world data collection. They reported distribution shifts in the dataset due to rotation transformation and orientation variation; as a result, model performance drops. Similarly, Ahen et al. Ahn et al. (2023) observed that changing the sensor location in Animal Activity Recognition reduces the model accuracy.

Najadat et al. trained an activity recognition model on smartphone-based data and tested the model against a dataset collected by smartwatches, causing a performance drop of 45% due to device and position variability Najadat et al. (2021). Khaked et al. reviewed the impact of subject, device, position, and orientation variability on the HAR task Khaked et al. (2025). They emphasized the significant effect of the subject variability, especially when performing complex activities, on the model performance. Additionally, they highlighted that sensor orientation hurts model accuracy less than sensor placement and type.

In recreational alpine skiing, any of the mentioned variability can happen, except that a solution requires a restricted sensor setup, e.g., full body sensor setup via Xsens MVN Technology Debertin et al. (2022), to control sensor and device-related shifts or a solution is developed for a particular group, e.g. only advanced and expert skiers using Connected Boots Neuwirth et al. (2020). However, a scalable sensor setup to target a broad range of skiers may rely on smartphones or smartwatches, which then add device variability. Furthermore, long-term recording of skiers' data using smartphones does not guarantee fixed orientation and placement since skiers locate their smartphones arbitrarily and may change their place and orientation several times during the recording. Azadi et. al. Azadi et al. (2025) have already investigated the effect of sensor placement and orientation and offered a preprocessing algorithm to mitigate this issue and ease continuous monitoring.

Additionally, recreational skiers perform skiing activities in various locations and seasons, which introduces environmental-related variability into the recorded skiing patterns, such as differences in skiing slopes and snow quality. Subject variability also significantly influences turning patterns, not only because each skier may execute a technique slightly differently, but also because subjects have different skill mastery Kranzinger et al. (2024); Azadi et al. (2022). However, it remains unclear how turning patterns change specifically as a function of skill level. It is reasonable to expect that less experienced skiers are unable to perform advanced techniques with biomechanical and behavioral consistency, introducing systematic variability into the dataset.

Studies and approaches in alpine skiing activity monitoring, Yoshioka et al. (2018); Supej and Holmberg (2021), and assessment, Federolf (2012); Procházka and Charvátová (2025), relying on wearable sensors especially inertial measurement units (IMU) are still limited to fixed sensor setups, Connected Boot sensor system Snyder et al. (2021); Kranzinger et al. (2024) or IMUs in combination with pressure sensors Matsumura et al. (2021), or several IMU sensors Pawlyta et al. (2019); Zhang et al. (2025), hindering them of being a scalable solutions proper for the wild nature of alpine skiing. A scalable solution, on the other hand, needs to handle real-world, long-term skiing sessions and recognize various skiing style patterns.

In this paper, we present a negative result: despite strong performance on benchmark and controlled subsets, Azadi et al. (2024), our models failed to generalize when deployed in the wild. We identify a previously underexplored cause, skier skill level, as the source of a significant distribution shift that undermines model performance. Although the skill level is not necessary for performing activities of daily living, it is crucial in sports, especially alpine skiing, which requires a combination of physical skills and technical proficiency. The current study discusses how the skill level of skiers can negatively impact skiing style recognition and model generalization.

2 Methodology

2.1 Dataset

The dataset used in this study consists of two parts. The first part is an alpine skiing dataset introduced in the Azadi et al. (2022). The dataset comprises IMU signals from skiers who performed various alpine skiing techniques selected by an expert based on the Austrian Ski Instructor Plan Österreichische Skischule (2021). These techniques are Parallel ski steering - long radii, Parallel ski steering - short radii, Dynamic parallel ski steering - long radii, Dynamic parallel ski steering - short radii, Carving - long radii, and Carving - short radii. We further adopted the method presented in Azadi et al. (2022) and conducted several data recordings without supervision, which optimally simulates a real-world use case, see Table 1.

The participants in the data collection possess varying skiing abilities, and we have categorized them into two groups based on their level of expertise: experienced and low-skilled. The experienced subjects include two alpine skiing instructors and five former ski racers; in total, seven experienced skiers coded as E in Table 1. The rest of the dataset contains ski enthusiasts who are familiar with all techniques and frequently go skiing during ski season; in total, eleven low-skilled skiers coded as LS in Table 1. This study was approved by the Ethics Committee of Johannes Kepler University with the protocol code JKU EC-24-2024.

The data was collected outside the laboratory in different skiing areas, slopes, and seasons in Austria. Therefore, the dataset has a high level of complexity, which reaches its maximum in the recordings without supervision. The IMU data (accelerometer, gyroscope, and magnetometer) were gathered using personal smartphones with a sampling rate of at least 50 Hz, and all recorded signals were resampled to 50 Hz, which contains sufficient information for high-frequency activities Yan et al. (2012). Table 1 summarizes the dataset used in this study, where sessions 6-9 were conducted without any supervision.

Session	Where	When	Subjects[1]	Skill[2]	Self-recorded	Glacier
1	Hintertux, Tyrol	June 2019	4	2*E,2*LS	No	Yes
2	Dachstein, Upper Austria	November 2019	2	ELS	No	Yes
3	Galterbergalm, Styria	February 2020	1	E	No	No
4	Hintertux, Tyrol	July 2020	3	E, 2*LS	Partially	Yes
5	Ramsau, Styria	February 2021	5	2*E, 3*LS	Yes	No
6	Obertauern, Salzburg	February 2021	3	3*E	Yes	No
7	Obertauern, Salzburg	March 2021	2	2*LS	Yes	No
8	Obertauern, Salzburg	December 2021	4	4*LS	Yes	No
9	Greifenburg, Carinthia	January 2022	1	E	Yes	No

Table 1: Data collection has been taken place in varied locations and conditions, including snow quality and slopes. Eighteen subjects recruited in this study have different capabilities ranging from novice to expert. However, we categorized them into two groups of experienced and low-skilled skiers. On the first three recordings, data collection was conducted using the provided smartphones (Galaxy S9). In the other recordings, data collection is done through the developed application on the subject's smartphone.
[1] Some of the skiers participated in the data collection more than one time.
[2] The abbreviation in the skill column is as follows: E: experienced, LS: low-skilled

We suggested that users attach their smartphones to their right side around their hip without any restriction on phone orientation. However, we anticipated that a fixed sensor placement and orientation would not be guaranteed in the recreational setting. Therefore, we relied on the motion analysis method, suggested by Azadi et al. (2025), to rotate recorded signals from any arbitrary position to a fixed reference frame as a required step before each learning task. The preprocessing algorithm fuses the accelerometer, gyroscope, and magnetometer using a two-step complementary filter and

then applies wavelet analysis to detect side motions, exhibiting turning behavior. The outcome of motion analysis delivers skiing motion in three axes: side, forward, and up.

2.2 Learning Task

We tackle the alpine skiing style recognition using various algorithms to investigate methodological issues when coping with a dataset, including different types of variability. The methods examined in this study involve traditional machine learning algorithms, which rely on hand-crafted domain-specific features, already investigated by Neuwirth et al. (2020), and autoencoder-based multi-task learning, proposed by Azadi et al. (2024). The multi-task learning architecture, formed of a multi-channel asymmetric autoencoder and a classification head for skiing style recognition, combines unsupervised (signal reconstruction) and supervised (skiing style recognition) tasks.

Moreover, we build a Variational Autoencoder(VAE), Kingma and Welling (2013), on top of the proposed autoencoder. VAEs offer several key advantages over traditional Autoencoders, primarily due to their probabilistic approach to learning the underlying probability distribution of the input data. Unlike standard autoencoders, which learn fixed latent representations, VAEs model the latent space as a probability distribution, which allows VAEs to handle uncertainty and variability in the data more effectively. However, they may perform poorly in reconstructing the input.

The models included an autoencoder-based multi-task learning model (AE-MTL), a variational autoencoder-based multi-task learning model (VAE-MTL), a single-task baseline (STL), and a classical Random Forest (RF) classifier, introduced in Figure 1. The diverse selection of models allowed for a comprehensive comparison between traditional machine learning methods and deep learning architectures capable of leveraging shared representations across multiple related tasks.

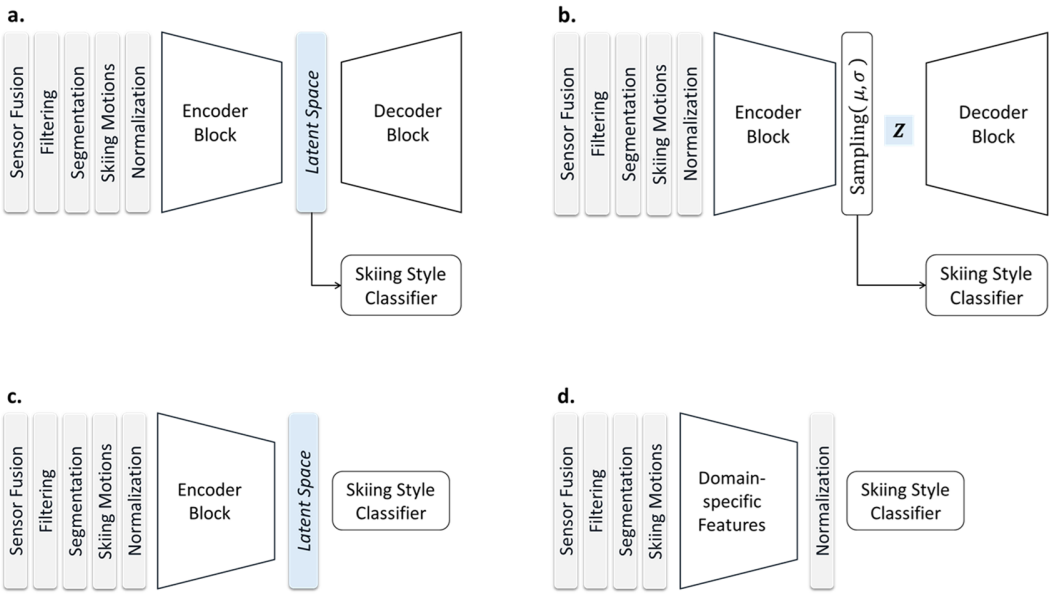

Figure 1: The figure displays an overview of the recognition models evaluated in this study for the style recognition task. a. autoencoder-based multi-task learning model (AE-MTL), b. variational autoencoder-based multi-task learning model (VAE-MTL), c. single-task baseline (STL), and d. classical Random Forest (RF) classifier

We developed AE-MTL and STL in the same way as introduced by Azadi et al. (2024), where the encoder dedicates a channel to each signal input to compress the signals into a latent space. The decoder in the AE-MTL, on the other hand, reconstructs the input signal in the output. The loss functions are cross-entropy, Li et al. (2020), for the classification tasks and Huber, Huber (1992), for the signal reconstruction. Additionally, we extended the AE-MTL by adding two dense layers to the encoder to form a Gaussian distribution characterized by a mean and a variance, thereby creating

VAE-MTL. The loss functions are cross-entropy for the classification tasks and KL divergence, Kullback and Leibler (1951), for the signal reconstruction.

The experiment uses Adam optimizer with a learning rate of 0.001 and the default hyperparameter and sets the batch size to 128 for all the analyses. The model was implemented using Tensorflow and Keras and ran on Intel(R) Core(TM) i7-7820HQ CPU @ 2.90GHz 2.90 GHz, Nvidia Quadro M2200, and 32 GB of installed RAM. An early stop was set to stop training when the validation loss stops improving by a minimum change of 0.0001 after 10 epochs.

To train the RF model, we extracted several domain-specific features from the skiing motions as the outcome of the preprocessing step. These features are turning dynamics, including turn cycle and maximum body inclination, and the range of movements in skiing motions. We employed a Random Forest Classifier using the Scikit-learn Python library, Pedregosa et al. (2011). The model was trained with 150 decision trees. All other hyperparameters were set to their default values. The same random_state was used for all runs to ensure reproducibility.

Finally, the study evaluates the introduced models against the entire dataset using Leave-One-Subject-Out Cross-Validation Bulling et al. (2014). Therefore, there is no overlap between training and test sets, i.e., a distinct user's recordings are used for model training and testing since a user may have several recordings.

3 Results

The overall model performance, as shown in Table 2, indicates poor generalization across all models. None of the models consistently achieved a high classification performance, meaning that generalization to unseen subjects remains a significant challenge in skiing style recognition based solely on IMU signals. While AE-MTL exhibited slightly higher scores, resulting in approximately 60% accuracy, the overall pattern suggests that the models struggled to adapt to the inter-subject variability inherent in recreational alpine skiing. The high standard deviation further illustrates the general difficulty of the task since all models exhibited considerable variability across subjects. This high spread suggests that subject-specific factors such as skiing style, execution quality, sensor placement variability, and skill level introduced substantial distribution shifts that the models failed to handle successfully.

Model	Accuracy	F1-score	Precision	Recall
Random Forest	54.10 ± 10.19	48.07 ± 13.62	52.41 ± 14.51	52.00 ± 12.35
Single-Task Learning	57.73 ± 14.86	48.41 ± 13.99	55.82 ± 15.65	53.02 ± 12.71
Multi-Task Learning	60.15 ± 13.60	51.10 ± 13.53	56.22 ± 15.38	55.69 ± 13.15
Variational Multi-Task Learning	57.60 ± 14.61	49.28 ± 14.14	55.32 ± 16.12	53.40 ± 13.58

Table 2: Overall performance of skiing style recognition models indicate a poor generalization to unseen skiers. Values ($mean \pm std$) for the f1-score, precision, and recall are macro average. Among models, AE-MTL performs slightly better, ensuing about 60% accuracy. However, all models exhibit high standard deviations, showing unstable recognition behavior.

Figure 2 illustrates classification models' performance per skiing style. Scanning the box plots shows that classification performance varied significantly across various skiing styles and exhibited an apparent drop for the Carving techniques (T5 and T6). All models displayed relatively higher difficulty distinguishing between biomechanically similar techniques, such as carving turns and dynamic parallel turns. For instance, the recall for carving - short radii (T6) is dramatically low, and simultaneously, the recall for dynamic parallel - short radii (T4) is greater than its precision, meaning models falsely classified other techniques as T4.

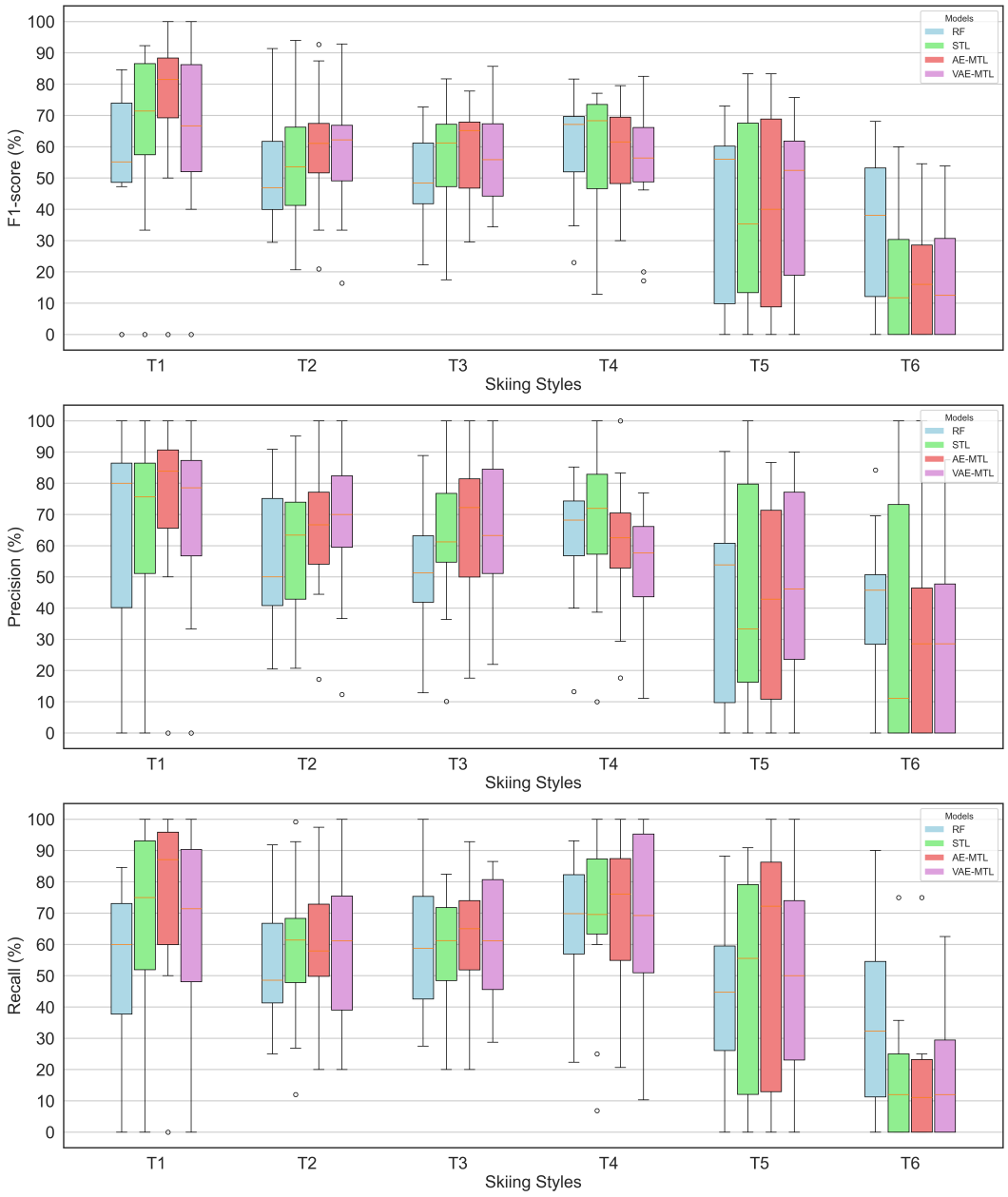

Figure 2: Classification metrics per technique shows a considerable performance drop on the Carving skiing styles. Since these techniques are the most advanced ones, they can be performed noticeably differently by low-skilled skiers.
∗ Abbreviations for techniques: T1: Parallel ski steering - long radii, T2: Parallel ski steering - short radii, T3: Dynamic parallel ski steering - long radii, T4: Dynamic parallel ski steering - short radii, T5: Carving - long radii, T6: Carving - short radii

To explain the poor generalization in style recognition, we further investigated skill-related variability in the dataset. For this reason, we reserved the recordings without supervision as test sets and trained models on the rest of the data. Then, we tested models against the unseen test sets categorized into experienced and low-skilled groups. The classification results in Table 3 indicate a significant difference between the two groups. Considering that all types of variability appears in the two test sets, the discrepancy can be related to the skill level of skiers.

Model	Accuracy		F1-score		Precision		Recall	
	LS	Exp.	LS	Exp.	LS	Exp.	LS	Exp.
Random Forest	33.94	58.46	22.12	52.66	27.48	62.15	24.90	51.14
Single-Task Learning	33.70	53.85	21.86	53.94	24.59	56.47	23.79	55.07
Multi-Task Learning	35.16	52.07	22.76	49.76	26.98	57.43	24.40	49.53
Variational Multi-Task Learning	35.90	50.30	27.13	50.97	30.14	59.38	31.23	50.32

Table 3: The classification outcomes show a clear difference between the two skill groups. The accuracy and macro-averaged metrics for the low-skilled group even falls in a random performance range. LS: low-skilled and Exp.: experienced

We further examined the AE-MTL model performance when tested against two subjects with varied skill levels who performed their activities in the same setting. The examination revealed a substantial discrepancy in model performance and latent space, Figure 3. Figure 3.a demonstrates confusion matrices in which the style recognition result for the experienced skiers is acceptable except on the carving - short radii (T6). On the other hand, although the recall for the first three techniques (T1-T3) from the low-skilled skier is within an acceptable range, the model could not recognize the more advanced techniques performed by the low-skilled skier, Figure 3.b.

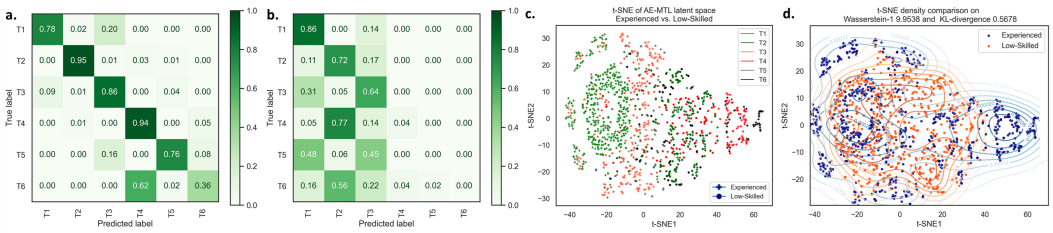

Figure 3: Skiing style classification for the experienced skier, (a.), although acceptable, illustrates significant confusion among carving and parallel dynamic turns, especially short radius turns. (b.) shows a noticeable performance drop and confusion for the low-skilled skier. (c.) and (d.) depict the latent space to compare features from the subjects, indicating a high Wasserstein-1 distance of 9.95, suggesting a distribution shift due to the skill level difference.
∗ Abbreviations for techniques: T1: Parallel ski steering - long radii, T2: Parallel ski steering - short radii, T3: Dynamic parallel ski steering - long radii, T4: Dynamic parallel ski steering - short radii, T5: Carving - long radii, T6: Carving - short radii

The latent space, Figure 3(c. and d.) visualized using t-SNE Maaten and Hinton (2008), explains the poor generalization, where latent features of short radii turning styles for the low-skilled skier are far from those for the experienced skier, and more in the center of the density. The Wasserstein-1 distance, Panaretos and Zemel (2019), between skill groups is 9.95, and the symmetric KL divergence, Zhang et al. (2023), is 0.56, suggesting a moderate distribution shift. The latent representations of experienced and low-skilled skiers are partially separable but overlapping, specifically, features from the T4 and T6 from the experienced skier on the right side. Figure 3, thus, indicates that skill level influences the feature distribution in a noticeable but not dominant way.

The sample in Figure 3.a, an experienced skier, achieved the highest accuracy. Classification metrics for this sample are 82% accuracy and 78%, 77%, and 76% macro average precision, recall, and f1-score, respectively. On the other hand, the same metrics for the low-skilled skier are 54%, 36%, 38%, and 32%, respectively, Figure 3.b. However, in the worst-case scenario, metrics can degrade to the level of random performance.

Figure 4 compares the density of latent features per technique. The result shows the Wasserstein-1 distance increases from left to right as styles become more challenging. Also, shorter turns exhibit a higher distance, where the density overlap is minimum, compared to their long radius version. The Wasserstein-1 distance of 5.39 and symmetric KL-divergence of 0.52 between experienced and low-skilled latent distributions indicate a statistically significant distribution shift of moderate magnitude for Parallel ski steering - long radii (T1): the two cohorts form separable but not entirely

isolated clusters in latent space. The separation reaches maximum for the carving - short radii (T6), where the latent features are fully dividable into two clusters, including several artifacts from the low-skilled skier, with Wasserstein-1 distance of 41.85 and symmetric KL-divergence of 10.31, demonstrating an evident and significant distribution shift.

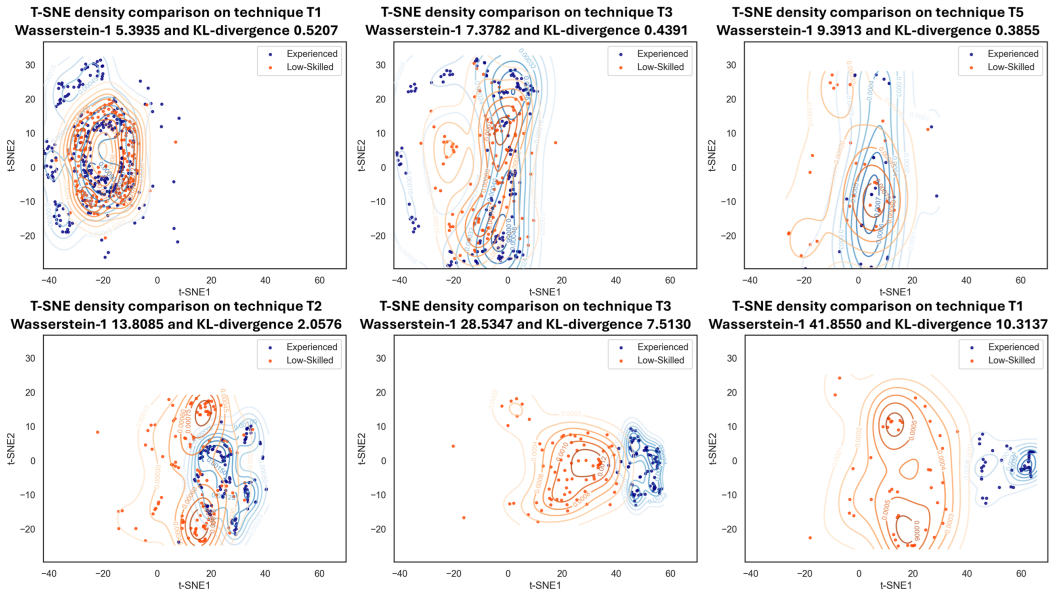

Figure 4: Density comparison per technique indicates a distribution shift as skiing style becomes more challenging, which is more critical for short turn activities. Wasserstein-1 distances are 5.39, 13.8, 7.37, 28.53, 9.39, and 41.85. The difference reaches the highest for the Carving - short radii, where there is now overlap between the two subjects.
∗ Abbreviations for techniques: T1: Parallel ski steering - long radii, T2: Parallel ski steering - short radii, T3: Dynamic parallel ski steering - long radii, T4: Dynamic parallel ski steering - short radii, T5: Carving - long radii, T6: Carving - short radii

4 Discussion

This paper explored skiing style recognition under real-world conditions and the complexity of the recognition task in recreational alpine skiing, where a wide range of variability exists, including sensor placement and orientation changes, device differences, location-specific factors, and, most importantly, skill level variability. The experiment showed that even well-established deep learning pipelines, here the AE-MTL model, could not generalize to real-world skiing data, even though they worked well on benchmark datasets. The main reason for this failure was the difference in skill levels among skiers. Our results showed that low-skilled and experienced skiers produce different movement patterns, which makes it harder for the model to recognize styles correctly. Yet, having an accurate style recognition pipeline is crucial when providing direct feedback on a skier's performance in a particular turning technique. The current research summarizes activity recognition issues as follows:

Some turning styles are very similar, e.g., parallel dynamic - short radii and carving - short radii. Even though an experienced skier can exhibit slightly different turning behavior using any of those two styles, any change in the skier's speed or body inclinations may make those techniques identical. The significant confusion between the carving and dynamic parallel turning techniques raises a question: Is the captured data on IMUs representative enough to distinguish these techniques? Or is attaching an IMU somewhere on the upper body suitable? Since having the sensor placed on the leg or boot may capture more harsh movements due to carvings.

Results show considerable recognition confusion among similar styles, which can be related to how skiers execute skiing turns. There are two primary explanations for why subjects perform activities differently; both are related to their skill levels. First, advanced skiers tend to adapt their style after

becoming masters at one level, which adds a few adjustments to their styles and, consequently, their patterns on the IMU signals. Second, low-skilled skiers struggle to execute advanced turning styles properly, and as a result, they may modify the style to a similar but less advanced skiing technique. We observed a significant impact of skier skill level on recorded IMU patterns, suggesting a clear distribution shift between low-skilled and experienced skiers. Further examination of the model performance, when testing against both skill groups, demonstrated a poor model generalization when facing data from the low-skilled group, Table 3. These results support the conclusion that skill level variability poses a persistent challenge for robust activity recognition.

Additionally, we measured the shift between distributions of latent features for experienced versus low-skilled skiers using the Wasserstein-1 distance and symmetric KL divergence. Because the Wasserstein-1 distance quantifies the minimum average displacement needed to transform one probability mass into the other, larger values directly indicate that the two groups occupy increasingly distinct regions of latent space. Also, the symmetric KL divergence quantifies how much the feature density of one skill group fails to explain the other. Low values (e.g., KL = 0.25, the minimum in Figure) indicate overlap and modest failure, while high values (KL > 5) reflect substantial distribution mismatch. In the carving-short condition, KL = 10.3 confirms that the low-skilled samples lie almost entirely outside the expert distribution, which is reduced to 8 after class combination, yet is high. Additionally, a high Wasserstein-1 distance was constantly observed, which, together with feature mass separations in the feature space, indicates the skill-related distribution shift.

Although deep-learning architectures such as the AE-MTL model achieve state-of-the-art accuracy on standard HAR benchmarks, this study shows that the same model suffers from poor generalization in a domain like recreational alpine skiing since domain-specific variability is broader than generic HAR variability. Recreational skiing layers sensor placement and orientation change, device heterogeneity, slope and snow conditions, mixed turn radii, and most critically skill level differences on top of the usual between-subject variation. Transition from a controlled lab setting to the real world using a living lab approach, therefore, requires explicitly accommodating the full spectrum of real-world variability that a domain introduces. This study has detailed the variability forms observed in recreational alpine skiing, and their impacts on the extent to which they can hurt the data, and offered solutions to address them, including the Skier Fixed Reference Frame.

Skill level damages both patterns and labels. Low-skilled participants often intend to, for instance, carve but produce dynamic parallel turns, or switch styles mid-run to regain control, which can also happen to experienced skiers but less often. This dual effect, although on its own an indication of skill level, implies that a model trained on clean experienced skiing patterns confronts out-of-distribution signals and ambiguous ground truth in skiing style recognition at deployment time. A central scientific insight is the identification and quantification of a skill-related distribution shift in recreational alpine skiing data. By measuring divergences in the latent representations (e.g., Wasserstein-1 distances up to 41.85 between experienced and low-skilled latent features), the study indicated that shifts in skier skill alone generate out-of-distribution signals and label noise that degrade skiing style recognition in comparison to generic HAR models. This finding underscores the need for future recognition systems to explicitly detect and adapt to skill-induced distributional changes. This shift helps explain the observed drop in style recognition accuracy when models are evaluated across diverse user groups.

All participants performed every technique in the same recording session with similar skis. However, carving turns are normally executed on specialized, short-radius carving skis. Because ski side-cut, stiffness, and length directly influence dynamics, the inertial patterns captured here combine technique with equipment constraints. It holds the same for skiing slopes since some of the recordings are conducted on an identical skiing slope regardless of the underlying skiing style, considering the first two techniques might fit the best to the blue slopes, parallel dynamic styles to the red, and carvings to black slopes.

The study reviews short and long radius turns, while in reality, skiers may perform techniques with medium radius turns, which are not introduced in the teaching plan. Skiers may execute such turns due to the skiing conditions, such as piste width or the crowd on the way. Furthermore, there are mixed activities that consist of more than one turning style, which are also due to the conditions on the slope, most likely to adjust the speed. Additionally, ground-truth labels in recreational skiing often reflect intended rather than actual execution, meaning that a subject may aim for one skiing style but ends up executing another technique. If a participant deviated from the instructed style, even briefly, the corresponding segment was still tagged with the intended label, introducing potential

annotation noise. All these conditions can introduce label noise that degrades supervised learning in the skiing style recognition.

5 Conclusion and Future Work

In this manuscript, we discussed the results of skiing style recognition in recreational alpine skiing and explained why the approach failed in the real world. The findings highlighted challenges that variability in skill level introduced into skiing style recognition. These observations underline the complexity of training fully generalizable recognition systems for recreational skiing. The result suggested that more experienced skiers generated more similar patterns. Although the models did not generalize well, they performed better when tested against more experienced skiers.

The investigation on skiing style recognition suggested a skill-related distribution shift in the dataset. A subject to investigate in the future is how the skill level of skiers (or users in other domains) affects the training. Nonetheless, two considerations warrant particular attention. First, the generated patterns for a particular technique vary among skiers holding various skill levels. Second, subjects with lower skill levels cannot perform advanced turning styles; therefore, the generated patterns might be random and invalid. This effect can be regarded as Out-of-Distribution (OOD) and should be addressed separately. For instance, an OOD detection method can be incorporated into the model training. Although we documented a clear skill-related distribution shift in recreational skiing data, the question remains as to how to mitigate the skill level impact on recognition models. Future work should explore domain-adaptation techniques, such as adversarial training Bai et al. (2021), to explicitly enforce invariance to skier skill level. By treating skill as a latent "domain", such approaches can encourage feature representations that remain stable across various skills.

6 Acknowledgment

This work has been supported by the FFG COMET K1 Center "Pro2Future II" (Cognitive and Sustainable Products and Production Systems of the Future), Contract No. 911655, by the FFG in the project "FedAI4Industry" Contract No. 921372, and by the provincial government of Upper Austria in the project "StreamingAI".

References

Ahn, S.-H., Kim, S., and Jeong, D.-H. (2023). Unsupervised domain adaptation for mitigating sensor variability and interspecies heterogeneity in animal activity recognition. *Animals*, 13(20):3276.

Azadi, B., Haslgrübler, M., Anzengruber-Tanase, B., Grünberger, S., and Ferscha, A. (2022). Alpine skiing activity recognition using smartphone's imus. *Sensors*, 22(15):5922.

Azadi, B., Haslgrübler, M., Anzengruber-Tanase, B., Sopidis, G., and Ferscha, A. (2024). Robust feature representation using multi-task learning for human activity recognition. *Sensors*, 24(2):681.

Azadi, B., Haslgrübler, M., and Ferscha, A. (2025). Motion analysis in alpine skiing: Sensor placement and orientation-invariant sensing. *Sensors*, 25(8):2582.

Bai, T., Luo, J., Zhao, J., Wen, B., and Wang, Q. (2021). Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*.

Bulling, A., Blanke, U., and Schiele, B. (2014). A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)*, 46(3):1–33.

Debertin, D., Wachholz, F., Mikut, R., and Federolf, P. (2022). Quantitative downhill skiing technique analysis according to ski instruction curricula: A proof-of-concept study applying principal component analysis on wearable sensor data. *Frontiers in bioengineering and biotechnology*, 10:1003619.

Federolf, P. A. (2012). Quantifying instantaneous performance in alpine ski racing. *Journal of sports sciences*, 30(10):1063–1068.

Gil-Martín, M., López-Iniesta, J., Fernández-Martínez, F., and San-Segundo, R. (2023). Reducing the impact of sensor orientation variability in human activity recognition using a consistent reference system. *Sensors*, 23(13):5845.

Hoelzemann, A., Romero, J. L., Bock, M., Laerhoven, K. V., and Lv, Q. (2023). Hang-time har: A benchmark dataset for basketball activity recognition using wrist-worn inertial sensors. *Sensors*, 23(13):5879.

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer.

Jimale, A. O. and Mohd Noor, M. H. (2023). Subject variability in sensor-based activity recognition. *Journal of Ambient Intelligence and Humanized Computing*, 14(4):3261–3274.

Khaked, A. A., Oishi, N., Roggen, D., and Lago, P. (2023). Investigating the effect of orientation variability in deep learning-based human activity recognition. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*, pages 480–485.

Khaked, A. A., Oishi, N., Roggen, D., and Lago, P. (2025). In shift and in variance: Assessing the robustness of har deep learning models against variability. *Sensors*, 25(2):430.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. (2021). Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR.

Kranzinger, C., Kranzinger, S., Hollauf, E., Rieser, H., and Stöggl, T. (2024). Skiing quality analysis of recreational skiers based on imu data and self-assessment. *Frontiers in Sports and Active Living*, 6:1495176.

Kreil, M., Sick, B., and Lukowicz, P. (2016). Coping with variability in motion based activity recognition. In *Proceedings of the 3rd International Workshop on Sensor-based Activity Recognition and Interaction*, pages 1–8.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Li, L., Doroslovački, M., and Loew, M. H. (2020). Approximating the gradient of cross-entropy loss function. *IEEE access*, 8:111626–111635.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Matsumura, S., Ohta, K., Yamamoto, S.-i., Koike, Y., and Kimura, T. (2021). Comfortable and convenient turning skill assessment for alpine skiers using imu and plantar pressure distribution sensors. *Sensors*, 21(3):834.

Min, C., Mathur, A., Montanari, A., and Kawsar, F. (2019). An early characterisation of wearing variability on motion signals for wearables. In *Proceedings of the 2019 ACM International Symposium on Wearable Computers*, pages 166–168.

Najadat, H., Ebrahim, M., Alsmirat, M., Shatnawi, O., Al-Rashdan, M. N., and Al-Aiad, A. (2021). Investigating the classification of human recognition on heterogeneous devices using recurrent neural networks. *Sustainable and Energy Efficient Computing Paradigms for Society*, pages 67–80.

Neuwirth, C., Snyder, C., Kremser, W., Brunauer, R., Holzer, H., and Stöggl, T. (2020). Classification of alpine skiing styles using gnss and inertial measurement units. *Sensors*, 20(15):4232.

Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6(1):405–431.

Pawlyta, M., Hermansa, M., Szczęsna, A., Janiak, M., and Wojciechowski, K. (2019). Deep recurrent neural networks for human activity recognition during skiing. In *International Conference on Man–Machine Interactions*, pages 136–145. Springer.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Procházka, A. and Charvátová, H. (2025). Wearable sensors and computational intelligence in alpine skiing analysis. *IEEE Access*.

Serpush, F., Menhaj, M. B., Masoumi, B., and Karasfi, B. (2022). Wearable sensor-based human activity recognition in the smart healthcare system. *Computational intelligence and neuroscience*, 2022(1):1391906.

Snyder, C., Martínez, A., Jahnel, R., Roe, J., and Stöggl, T. (2021). Connected skiing: Motion quality quantification in alpine skiing. *Sensors*, 21(11):3779.

Sopidis, G., Haslgrübler, M., and Ferscha, A. (2023). Counting activities using weakly labeled raw acceleration data: A variable-length sequence approach with deep learning to maintain event duration flexibility. *Sensors*, 23(11):5057.

Supej, M. and Holmberg, H.-C. (2021). Monitoring the performance of alpine skiers with inertial motion units: practical and methodological considerations. *Journal of Science in Sport and Exercise*, 3(3):249–256.

Yan, Z., Subbaraju, V., Chakraborty, D., Misra, A., and Aberer, K. (2012). Energy-efficient continuous activity recognition on mobile phones: An activity-adaptive approach. In *2012 16th international symposium on wearable computers*, pages 17–24. Ieee.

Yoshioka, S., Fujita, Z., Hay, D. C., and Ishige, Y. (2018). Pose tracking with rate gyroscopes in alpine skiing. *Sports Engineering*, 21:177–188.

Zhang, Y., Fei, Q., Chen, Z., and Liu, X. (2025). A temporal-channel-spatial attention network for skiing recognition based on multi-graph generation. *IEEE Transactions on Instrumentation and Measurement*.

Zhang, Y., Pan, J., Li, L. K., Liu, W., Chen, Z., Liu, X., and Wang, J. (2023). On the properties of kullback-leibler divergence between multivariate gaussian distributions. *Advances in neural information processing systems*, 36:58152–58165.

Österreichische Skischule, D. (2021). *Snowsport Austria*. Publisher Brothers Hollinek I& Co. GmbH, 3 edition.