Machine Learning Architectures to Classify Activities of Daily Living and Fall Types From Wearable Accelerometer Data

A. Antonietti

Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milano, Italy
alberto.antonietti@polimi.it

***Abstract*— This study compares different machine learning (ML) approaches to classify motion-based activities of daily living based on accelerometer measurements, focusing on identifying different types of falls. Falls are a significant cause of injuries and deaths in older adults, emphasizing the need for real-time fall detection and alert systems. ML algorithms have shown high accuracy in detecting falls in experimental settings, but their performance in real-world scenarios still needs to be studied.**

Two publicly available datasets, collected from healthy subjects with smartphone accelerometers, were used for evaluation. Different ML classifiers (Support Vector Machine, Random Forest, CatBoost, and a meta-algorithm) were tested using raw accelerometer data and selected features in the time and frequency domains. Results demonstrated that CatBoost and Random Forest outperformed Support Vector Machine in both datasets when using a carefully chosen set of features instead of the raw data as input. CatBoost, although accurate, showed higher computational costs, making Random Forest a more practical choice for real-world applications. The meta-ML system did not provide a significant advantage over individual algorithms.

This research contributes insights into feature selection and computational efficiency in accelerometer data classification. It provides evidence-based recommendations for practitioners working with activities of daily living classification and fall detection, aiming to enhance the use of ML in real-world scenarios and improve the quality of life for individuals requiring monitoring and fall prevention.

Keywords*— *Inertial sensors, IMU, Support Vector Machine, Random Forest, CatBoost, Generalizability.

I. Introduction

The ability to independently perform Activities of Daily Living (ADLs) is crucial for maintaining autonomy and high quality of life. ADLs encompass fundamental self-care tasks and different types of movements, including walking at various paces, running, going up and down the stairs, etc. As the global population ages and the prevalence of chronic conditions rises, there is a growing interest in leveraging advanced technologies to monitor and assist individuals in their daily activities.

In recent years, the use of machine learning algorithms in combination with wearable accelerometers has emerged as a promising approach for classifying motion-based ADLs [1–3]. Accelerometers are small, unobtrusive devices capable of capturing the body's movements, making them ideal for real-time activity monitoring. Machine learning techniques enable the extraction of meaningful patterns and features from accelerometer data, facilitating the identification and classification of specific ADLs.

The classification of distinct ADLs based on accelerometer measurements has gained significant attention due to its potential applications in real-time activity tracking and diagnostics of unwanted events, particularly falls. Machine learning algorithms have been widely used in this domain to accurately detect and classify activities. Previous studies have investigated the optimal placement of accelerometers on the body for activity detection [4]. The hip has been found to be the best single location for recording data, providing better accuracy than other locations [4]. Additionally, combining data from multiple sensors has been shown to improve classification accuracy [5]. However, it is important to note that the accuracy of classification algorithms trained in laboratory conditions may decrease when applied to free-living subjects [3, 6]. Therefore, further research is needed to evaluate the reproducibility of classification algorithms in daily life settings. This paper aims to contribute to the existing literature by proposing a comparison between different machine learning approaches to classify activities of daily living based on accelerometer measurements, with a focus on identifying different types of falls.

The motivation behind this research is to keep track of users' activities in real-time and provide possible diagnostics of unwanted and unexpected events, with special attention to movements during human falling and the distinction among various types of falls. Falls are a major cause of injuries and deaths in older adults, and even when no injury occurs, individuals who fall often require assistance to get up [7]. Machine learning algorithms have been utilized to achieve high accuracy in detecting falls in experimental settings [8, 9]. However, studies examining the accuracy of fall detection algorithms in real-world settings are less common [7, 10]. Therefore, this research aims to validate the accuracy of machine learning-based fall detection systems using real-world fall and non-fall datasets.

II. Materials and Methods

We tested and compared four different machine learning classifiers, using either raw accelerometer data ($x-y-z$ axis) or 94 important features (in the time and frequency domains) as input. We trained and tested on a first

smaller and simpler ADLs dataset and then verified the generalizability of our findings on a larger and more complex dataset. Here, we describe in detail the datasets used, the two types of input data, the machine learning architectures, the statistical analysis performed, and a summary of the hardware and software used.

ADL Datasets

Both datasets have been collected on healthy subjects with a smartphone or smartphone-like accelerometers. The first dataset is called UniZgFall and contains data from 16 young healthy subjects performing 10 types of ADLs and simulated falls while wearing an inertial sensor unit (Shimmer sensing, Ireland) attached sideways to their waist [11]. Each item in the dataset is a $N \times 3$ matrix, where N is a variable number of samples, and each column contains accelerations (m/s^2) along the $x-y-z$ axes, sampled at 200 Hz. Each item is labeled with one of the following classes: Walking (MW), Running (MR), Jumping (MJ), Walking down the stairs (WD), Walking up the stairs (WU), Forward fall (FF), Sideways fall (FS), Backward fall (FB), Lying down (LD), Other tasks/inactive (OT). The dataset contains 468 items and is unbalanced: 68 items each for MW, MR, and MK, 13 items each for WD and WU, 34 items for FF, FS, FB, and LD, and 102 items for OT.

The second dataset is larger and more complex; it is called UniMiB SHAR and includes 1980 items of both human activities and falls performed by 30 subjects of ages ranging from 18 to 60 years [12]. The smartphone used in the experiments was a Samsung Galaxy Nexus I9250. Here too, each item is a $N \times 3$ matrix, but the sampling rate was set to 50 Hz. Items belong to 17 different classes as reported in Table 1. The dataset is unbalanced since physiological activities (#1-9) have 60 items each, while all fall types (#10-17) have 180 items.

Table 1. ADLs and fall types categories in UniMiB SHAR dataset.

#	Description	Label
1	From laying on the bed to standing	StandingUpFL
2	From standing to lying on a bed	LyingDownFS
3	From standing to sitting on a chair	StandingUpFS
4	Moderate running	Running
5	From standing to sitting on a chair	SittingDown
6	Climb the stairs moderately	GoingDownS
7	Down the stairs moderately	GoingUpS
8	Normal walking	Walking
9	Continuous jumping	Jumping
10	Fall backward while trying to sit on a chair	FallingBackSC
11	Generic fall backward from standing	FallingBack
12	Falls using strategies to prevent the impact	FallingWithPS
13	Fall forward from standing	FallingForw
14	Fall right from standing	FallingLeft
15	Fall right from standing	FallingRight
16	Falls with contact to an obstacle	HittingObstacle
17	Getting unconscious	Syncope

Input data

We wanted to test if the machine learning architecture led to higher accuracies when working with raw accelerometer data or with selected features in both the time and frequency domains. For this reason, two types of inputs were provided.

Raw data. In this case, for each item of the dataset, we built $N \times 3$ matrices containing the raw accelerometer signals measured along the 3 axes. Since we had to provide input data with a fixed shape, while each item has a variable number of samples, we set the number N of rows to the maximum length of the whole dataset, namely 5150 samples (i.e., 25.7 s) and 3208 (i.e., 64.2 s) for UniZgFall and UniMiB SHAR datasets, respectively. We performed a zero-padding for all items to transform them to the maximum shape.

Features. We selected and computed for each item 94 salient features: 50 in the time domain (16 for each axis, plus 2 combining more axis) and 44 on the computed Fast-Fourier transform (FFT) (14 for each axis, plus 2 combining more axis), as reported in Table 2. The majority of these features are intuitive and easy to understand. To compute the *Energy* of a signal in each axis, we take the mean of the sum of squares of the values within a window along that specific axis. The *Average resultant acceleration* is calculated by finding the average of the square roots of the squared values from each of the three axes, which are then added together. *Signal Magnitude Area* is defined as the average of the absolute values of the three axes. The 94 features used in our analysis were carefully selected based on comprehensive literature studies focusing on feature selection within $x-y-z$ accelerometer datasets used for classification purposes. Incorporating features from both the time and frequency domain is crucial for a comprehensive analysis of accelerometer data. Time domain features offer interpretability, capturing transient events and providing real-time relevance. Meanwhile, frequency domain analysis decomposes signals into constituent frequencies, revealing hidden patterns, aiding in noise filtering, and discerning subtle activities. Integrating both domains provides a complementary set of information that is supposed to enhance the classification capabilities of the algorithms.

Machine Learning Algorithms

Among the myriad of machine learning methods for the classification of accelerometer data, three methods have been selected: Support Vector Machine (SVM), Random Forest (RF), and CatBoost (CAT) [13]. SVM was chosen for its high degree of explainability and lightweight [14]. RF was selected because it is an effective method for ranking the importance of variables in a classification problem [15]. CatBoost, on the other hand, is a high-performance library for gradient boosting on decision trees that has shown superior performance in other domains, such as audio signals [16].

SVM is a well-established machine-learning method for classification problems. It utilizes a decision sur-

Table 2. Features extracted from the raw accelerometer signals

#	Description	Acceleration	FFT
1	Mean	✓	✓
2	Standard deviation	✓	✓
3	Average absolute deviation	✓	✓
4	Minimum	✓	✓
5	Maximum	✓	✓
6	Maximum - minimum	✓	✓
7	Median	✓	✓
8	Median absolute deviation	✓	✓
9	Interquartile range	✓	✓
10	Negative values count	✓	✓
11	Positive values count	✓	
12	Number of values above mean	✓	
13	Number of peaks	✓	✓
14	Skewness	✓	✓
15	Kurtosis	✓	✓
16	Energy	✓	✓
17	Average resultant acceleration	✓(1 value)	✓(1 value)
18	Signal magnitude area	✓(1 value)	✓(1 value)

face constructed in high-dimensional feature space. The support-vector network, which is the basis of SVM, has been shown to have high generalization ability and can handle non-separable training data [13, 14]. We used it with enabled probability estimates.

RF is a popular machine learning method that utilizes an ensemble of decision trees to make predictions. It is known for its ability to rank the importance of variables in a classification problem and has been used successfully in various domains, including physical activity classification using accelerometer data [5, 15]. We set the number of decision trees in the forest to 100.

CatBoost is a relatively new gradient boosting toolkit that has gained attention for its superior performance in various datasets. It incorporates innovative algorithms for processing categorical features and implements ordered boosting, a permutation-driven alternative to the classic boosting algorithm [16]. CAT has been shown to outperform other publicly available boosting implementations in terms of quality.

SVM, RF, and CAT are three machine-learning methods that have been selected for the classification of accelerometer data. SVM is chosen for its explainability and lightweight, RF for its ability to rank variable importance, and CatBoost for its superior performance in various domains. These methods have been widely used and have shown promising results in the classification of accelerometer data. In addition, we tried a multi-architecture decision system (called ALL) where the prediction of the testing classes was performed considering, for each item of the testing set, the class predicted with the highest confidence by SVM, RF, and CAT. If not differently stated, we used the default parameters for the three architectures in all cases.

In our study, we utilized a stratified 10-fold cross-validation approach to rigorously assess the performance and robustness of our machine learning model. Stratified cross-validation tackles this issue by ensuring that each fold maintains the same class distribution as the original dataset. By doing so, we obtain a more reliable estimate of the model's performance, especially when dealing with imbalanced datasets. This approach enhances the generalization capability of our models and provides a more robust assessment of their effectiveness across various class distributions. When creating the fold, we enabled shuffling.

Statistical Analysis

We used two metrics to evaluate the performance of the systems for the multi-class classification problem: accuracy (i.e., the number of correctly classified items over the number of test items) and the F1-score, computed through a multi-class formulation: each class against the others. For each class i, the F1-score is computed as:

$$F1-score_i = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (1)$$

The global score is defined as the average value between all the $F1-scores_i$. Since the time taken for the training is another important metric, we computed the training time needed by each fold. For RF and CAT, we could also analyze the relative importance of each feature. Finally, we visualize the per-class performance via normalized confusion matrices (averaging across the 10 folds). For each condition, we reported the metrics as median [25^{th} - 75^{th} percentiles]. We performed statistical tests to verify, in each condition, differences between the performance of SVM, RF, CAT, and ALL. Given the low sample size, we used non-parametric statistics: the Kruskal-Wallis test and post hoc analysis via Dunn's test with Bonferroni correction.

Hardware and Software

Data processing and training were performed on a virtual machine running Ubuntu 22.04.1 LTS, with *Python* 3.10.6, *sklearn* 1.2.2, *catboost* 1.1.1, *scipy* 1.10.1, *scikit posthocs* 0.7.0. The virtual machine had the following technical specs: Intel(R) Xeon(R) CPU E5-2690 v3 @2.60 GHz, 16 CPUs, 64 GB RAM memory. We did not use GPU accelerators. To guarantee a maximum degree of reproducibility, all the code for training and analysis is openly available at https://github.com/alberto-antonietti/ml_accelerometers.

III. Results

UniZgFall Dataset

All four algorithms did not work particularly well when the raw acceleration data were used as input (Figure 1) Left panels), with accuracies for SVM = 0.78 [0.72 0.83], for RF = 0.72 [0.68 0.79], CAT = 0.77 [0.70 0.81], ALL = 0.79 [0.70 0.87], and no significant differences (Kruskal-Wallis test p-value = 0.42). F1-scores for SVM = 0.67 [0.58 0.76], for RF = 0.66 [0.63 0.75], for CAT = 0.65 [0.60 0.74], ALL = 0.68 [0.59 0.82], without any difference (p-value: 0.96). RF was

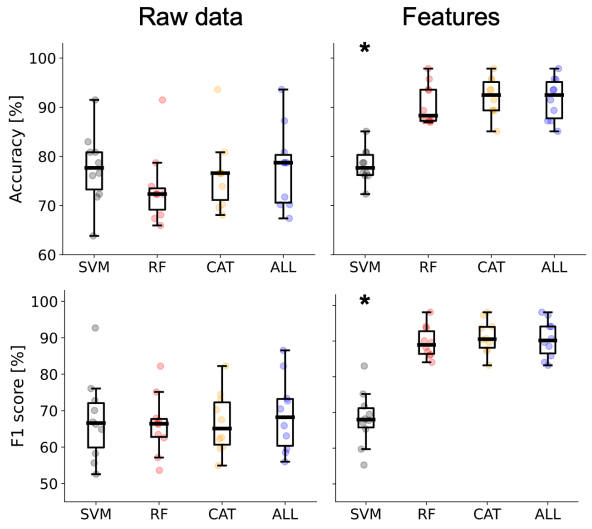

Figure 1. UniZgFall dataset. Accuracy and F1-scores boxplots for the four machine learning systems providing raw data (left panels) or selected features (right panels) as input .

very fast, 1.79 [1.59 2.08] s for training, SVM slower, 227.40 [196.47 256.27] s, and CAT extremely heavy 1456.68 [1445.38 1624.03] s. We do not report the time needed by ALL since it is given by the sum of SVM, RF, and CAT times.

When using the selected features as input, the performances definitively improved for all algorithms, but SVM (Figure 1) Right panels). Accuracies are SVM = 0.78 [0.76 0.81], RF = 0.88 [0.87 0.96], CAT = 0.92 [0.89 0.96], ALL = 0.92 [0.87 0.96]. Only SVM is significantly different from the other groups (Kruskal-Wallis test p-value: $5.6 \cdot 10^{-5}$). F1-scores are SVM = 0.68 [0.65 0.75], RF = 0.89 [0.86 0.94], CAT = 0.91 [0.88 0.97], ALL = 0.90 [0.86 0.97]. SVM is the only different one, too (p-value: $5.5 \cdot 10^{-5}$). Given the smaller size of the input, due to the feature selection, training times drastically decreased: SVM = 0.11 [0.11 0.11] s, RF = 0.42 [0.41 0.45] s, CAT 17.71 [17.43 17.90] s.

Analyzing in more detail the output of the best algorithm (CAT), the normalized confusion matrices (Figure 2) show a large number of misclassifications when using raw data and very good behavior when using features. In this case, CAT had some difficulty in distinguishing forward falls (FF) from the other two types of falls and between laying down (LD) from other unclassified tasks (OT). The most important feature for both RF and CAT was *Energy* along the y axis (3.5% and 5.0%, respectively). Features of the y axis were significantly more important than the other two axes (40.3% and 43.6%), as well as frequency features (58.6% and 51.2%).

UniMiB SHAR Dataset

We observed similar trends on the other dataset, which is larger (1980 items instead of 468) and more complex (17 classes instead of 10). Accuracies and F1-scores

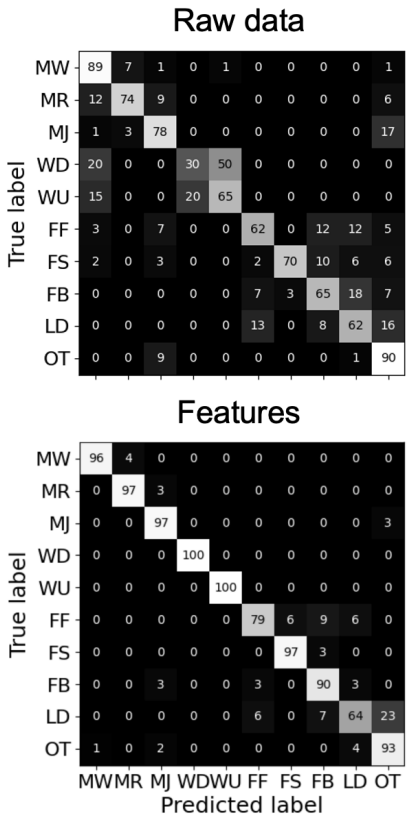

Figure 2. UniZgFall dataset. Normalized confusion matrix made averaging the results from the 10 folds with the CAT algorithm, providing as input raw data (upper panel) or selected features (lower panel).

were significantly better when using features as input, except for SVM, which had bad performances in both cases (Figure 3). More precisely, for raw data accuracies: SVM = 0.49 [0.44 0.53], RF = 0.56 [0.51 0.59], CAT = 0.55 [0.54 0.61], and ALL 0.55 [0.55 0.60], with SVM performing statistically lower than CAT and ALL (p-value = 0.0015). Similar output for the F1-score: SVM = 0.51 [0.48 0.55], RF = 0.57 [0.55 0.62], CAT = 0.60 [0.55 0.64], ALL = 0.60 [0.55 0.62] (p-value = 0.0019). Training times are in line with the results obtained on the UniZgFall dataset.

With features, we had the following accuracies: SVM = 0.32 [0.31 0.35], RF = 0.67 [0.63 0.71], CAT = 0.71 [0.68 0.73], ALL = 0.70 [0.69 0.74], and F1-scores: SVM = 0.31 [0.27 0.32], RF = 0.72 [0.68 0.75], CAT = 0.75 [0.71 0.77], ALL 0.75 [0.72 0.79]. In both cases, SVM performed significantly worse than the other three algorithms. SVM and RF continued to be pretty fast (2.23 [2.20 2.36] s and 1.71 [1.71 1.74] s, respectively), while CAT is computationally costly (35.76 [35.69 35.97] s).

The normalized confusion matrices (Figure 4) highlight where the CAT algorithm made the most errors. As expected, the results with raw data contain more spread

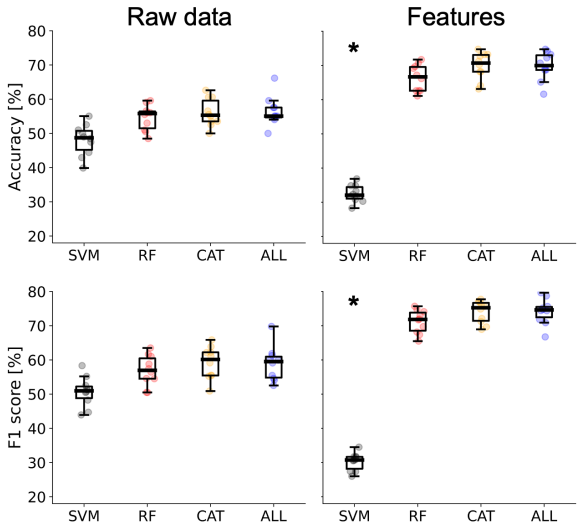

Figure 3. UniMiB SHAR dataset. Accuracy and F1-scores boxplots for the four machine learning systems providing as input raw data (left panels) or selected features (right panels).

misclassified items, while with features, CAT found it difficult to choose the correct fall type and to identify movements from and to horizontal positions (e.g., standing up after lying on a bed). Among the most important features, we found *above_mean* along the y and z axes for both RF and CAT. Features of the y and z axes were slightly more important (values around 34%) than the x axis (around 28%), and frequency features were less important (47.7% and 41.1% for RF and CAT, respectively) than time-domain ones.

IV. Discussion and Conclusions

In this paper, we presented a comprehensive analysis of multi-class accelerometer data classification using machine learning algorithms. Our study aimed to elucidate the key factors influencing model performance and generalizability, shedding light on the most effective approaches for real-world applications.

One of the fundamental aspects of our research is the utilization of two distinct datasets, which allowed us to verify the robustness and generalizability of our findings. Notably, the performances and trends obtained from both datasets were strikingly similar, indicating that our models can effectively generalize across different data sources [17]. It is essential to note that the consistency in performance across datasets highlights the robustness of the model type and architecture. However, it is imperative to recognize that while the fundamental model architecture remains consistent, the optimal parameters vary for each dataset. This nuanced understanding underscores the model's adaptability while emphasizing the importance of fine-tuning parameters to achieve optimal results for specific datasets.

We strategically selected a diverse set of features in both the time and frequency domains to address the computational load associated with using raw accelerometer data with long time series (i.e., up to tens of seconds). Our findings demonstrate that these carefully chosen features capture the salient elements of the 3-axes acceleration data, enabling accurate classification without sacrificing computational efficiency [18, 19].

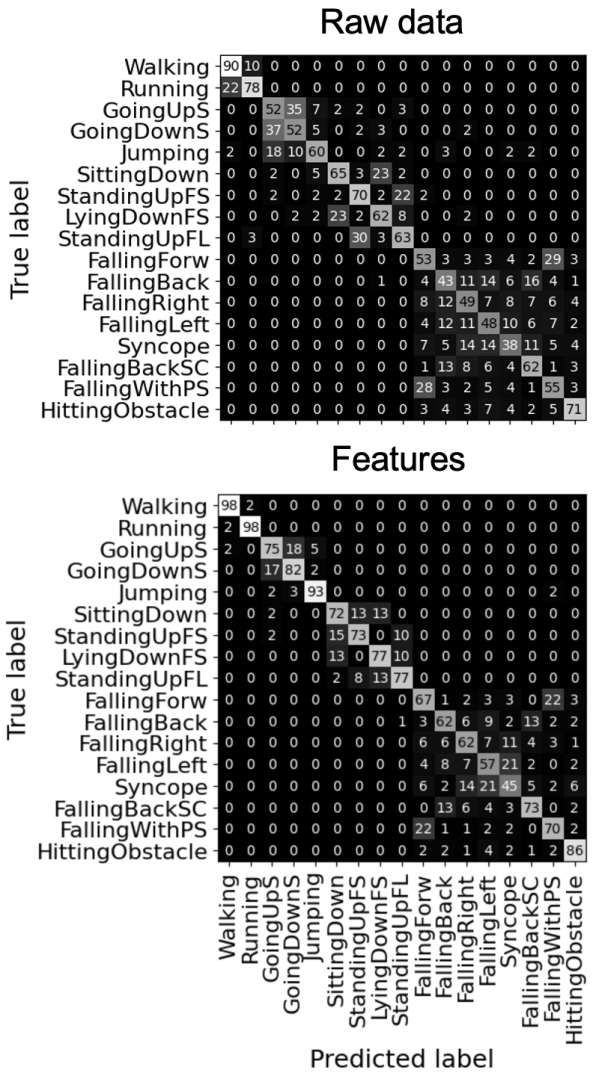

Figure 4. UniMiB SHAR dataset. Normalized confusion matrix made averaging the results from the 10 folds with the CAT algorithm, providing as input raw data (upper panel) or selected features (lower panel).

Among the machine learning algorithms evaluated, RF and CAT emerged as the top performers, yielding accuracies and F1-scores that align with single-test values reported in prior studies [20]. While the simplicity and widespread use of SVM in past studies were notable, our experiments revealed its relative ineffectiveness in this specific multi-class classification with accelerometer data. Although CAT showed a slight improvement in accuracy compared to RF, it came at a significant computational cost, being more than 40 times slower. Our

recommendation leans toward leveraging RF algorithms for real-world and real-time applications due to their balanced trade-off between accuracy and efficiency. However, we acknowledge that the performance of CAT, optimized for GPUs, warrants re-evaluation with GPU accelerators to ascertain its true potential [16]. Additionally, we explored the integration of a meta-machine learning system (ALL), combining the outputs of RF, SVM, and CAT, leveraging the confidence levels of each algorithm. Surprisingly, this approach did not yield any significant advantage, suggesting that the individual strengths of RF and CAT were sufficient for the multi-class classification task at hand.

The two datasets included in our analysis involved wearing accelerometers in specific locations close to the waist; this poses practical challenges, potentially leading to low user compliance. Particularly for the elderly, consistent placement of devices in specific directions during daily activities can be difficult. This limitation is common in cellphone app-based studies. Simple devices like necklaces with alarm buttons might offer operational advantages for effective fall detection due to their ease of use and user-friendliness.

In conclusion, our research highlights the importance of feature selection and the careful consideration of computational efficiency in multi-class accelerometer data classification. We contribute valuable insights into the performance of machine learning algorithms, providing practitioners with evidence-based recommendations for real-world applications.

ACKNOWLEDGEMENTS

Alberto Antonietti is funded by the Project "EBRAINS-Italy (European Brain ReseArch INfrastructureS-Italy)" granted by European Union – NextGenerationEU adopted by the Italian Ministry of University and Research, CUP B51E22000150006.

REFERENCES

[1] I. M. Pires, N. M. Garcia, E. Zdravevski, and P. Lameski, "Activities of daily living with motion: A dataset with accelerometer, magnetometer and gyroscope data from mobile devices," *Data in Brief*, vol. 33, p. 106628, Dec. 2020.

[2] M. Janidarmian, A. Roshan Fekr, K. Radecka, and Z. Zilic, "A Comprehensive Analysis on Wearable Acceleration Sensors in Human Activity Recognition," *Sensors*, vol. 17, p. 529, Mar. 2017.

[3] F. Bagalà, C. Becker, A. Cappello, L. Chiari, K. Aminian, J. M. Hausdorff, W. Zijlstra, and J. Klenk, "Evaluation of Accelerometer-Based Fall Detection Algorithms on Real-World Falls," *PLoS ONE*, vol. 7, p. e37062, May 2012.

[4] I. Cleland, B. Kikhia, C. D. Nugent, A. Boytsov, J. Hallberg, K. Synnes, S. McClean, and D. D. Finlay, "Optimal Placement of Accelerometers for the Detection of Everyday Activities," *Sensors*, 2013.

[5] K. V. Ellis, J. Kerr, S. Godbole, J. Staudenmayer, and G. R. G. Lanckriet, "Hip and Wrist Accelerometer Algorithms for Free-Living Behavior Classification," *Medicine & Science in Sports & Exercise*, 2016.

[6] Gyllensten, Illapha Gustav Lars Cuba and A. G. Bonomi, "Identifying Types of Physical Activity With a Single Accelerometer: Evaluating Laboratory-Trained Algorithms in Daily Life," *Ieee Transactions on Biomedical Engineering*, 2011.

[7] O. Aziz, J. Klenk, L. Schwickert, L. Chiari, C. Becker, E. J. Park, G. Mori, and S. N. Robinovitch, "Validation of Accuracy of SVM-based Fall Detection System Using Real-World Fall and Non-Fall Datasets," *Plos One*, 2017.

[8] K.-C. Liu, K.-H. Hung, C.-Y. Hsieh, H.-Y. Huang, C.-T. Chan, and Y. Tsao, "Deep-Learning-Based Signal Enhancement of Low-Resolution Accelerometer for Fall Detection Systems," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, pp. 1270–1281, Sept. 2022.

[9] S. R. M. Edeib, R. A. Dziyauddin, and N. I. M. Amir, "Fall Detection and Monitoring using Machine Learning: A Comparative Study," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 2, 2023.

[10] J. A. Santoyo-Ramón, E. Casilari-Pérez, and J. M. Cano-García, "A study on the impact of the users' characteristics on the performance of wearable fall detection systems," *Scientific Reports*, vol. 11, p. 23011, Nov. 2021.

[11] D. Razum, G. Seketa, J. Vugrin, and I. Lackovic, "Optimal threshold selection for threshold-based fall detection algorithms with multiple features," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, (Opatija), pp. 1513–1516, IEEE, May 2018.

[12] D. Micucci, M. Mobilio, and P. Napoletano, "UniMiB SHAR: A Dataset for Human Activity Recognition Using Acceleration Data from Smartphones," *Applied Sciences*, vol. 7, p. 1101, Oct. 2017.

[13] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, 1995.

[14] J. Alizadeh, M. Bogdan, J. Classen, and C. Fricke, "Support Vector Machine Classifiers Show High Generalizability in Automatic Fall Detection in Older Adults," *Sensors*, vol. 21, p. 7166, Oct. 2021.

[15] K. Ellis, J. Kerr, S. Godbole, G. Lanckriet, D. Wing, and S. Marshall, "A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers," *Physiological Measurement*, vol. 35, pp. 2191–2203, Dec. 2014.

[16] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, (Red Hook, NY, USA), pp. 6639–6649, Curran Associates Inc., Dec. 2018.

[17] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SIGKDD Explorations Newsletter*, vol. 12, pp. 74–82, Mar. 2011.

[18] C. P. Burgos, L. Gartner, M. A. G. Ballester, J. Noailly, F. Stocker, M. Schonfelder, T. Adams, and S. Tassani, "In-Ear Accelerometer-Based Sensor for Gait Classification," *IEEE Sensors Journal*, vol. 20, pp. 12895–12902, Nov. 2020.

[19] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, pp. 107–115, Mar. 2021.

[20] A. H. K. Montoye, B. S. Westgate, M. R. Fonley, and K. A. Pfeiffer, "Cross-validation and out-of-sample testing of physical activity intensity predictions with a wrist-worn accelerometer," *Journal of Applied Physiology*, vol. 124, pp. 1284–1293, May 2018.