Human Activity Recognition for Complex Soldier Movements

Andrew Tweedell[1]

[1]Khoury College of Computer Science
Northeastern University, Boston MA

Abstract

The military intends to deploy many Soldier-borne technologies that depend on an accurate and robust identification of Soldier physical state. Human Activity Recognition (HAR) methods aimed at classifying human behavior for military applications need to account for unique Soldier movements not present in typical open source data sets. The purpose of the project was to derive and evaluate classification methods for Soldier relevant tasks from inertial measurement units (IMUs). The tasks included dynamic movements like crawling, wall climb and rushing. Using time-domain statistics of biomechanical features from 14 IMU devices situated on different limbs, Random Forest (RF) Classifier, Support Vector Machines (SVM), and Gradient Boosted Classifier (GBC) models were trained and evaluated. Additionally, feature reduction techniques using random forest selection and principal components analysis (PCA) were evaluated as potential pre-processing approaches for more efficient computation. The RF model with the full feature set produced the highest test F1 score (0.99) with only minimal increases in training time beyond SVM methods. Both SVM and GBC also produced high test F1 scores (0.93 and 0.92, respectively), while the SVM provided the fastest fit time (0.05 sec). PCA dimensionality reduction produced the worst generalization ability, decreasing the test F1 scores below 0.66. Gradient Boosting requires extensive training time making it intractable for real-world applications. Using IMU sensors and biomechanical features, we can achieve excellent performance when classifying unique Soldier physical tasks.

Introduction

Many technologies the military intends to deploy for Soldiers rely on being able to classify the physical behavior of the human at any point in time. Human activity recognition (HAR) is a machine learning approach for predicting and labeling a specific movement or action a human is performing based on sensor data. Indeed, HAR has been an active area of research for decades; however, sensor technology has only recently gotten small enough and processing units gotten efficient enough that models could be developed off data collected and subsequently deployed for many real-world applications, including health care (Bao and Intille 2004) , sports (Ladh et al. 2013), and human-technology interaction (Tapia, Intille, and Larson 2004).

The types of sensor modalities used and the type of models developed for HAR are very diverse and depend primarily on the type of activity and the complexity of the feature space (Bulling, Blanke, and Schiele 2014). Regardless of specific modality, each data type represents time series signals. Hand-crafted signal features derived from wearable sensor signals are strong candidates for traditional machine learning algorithms, such as random forest (RF) and logistic regression, whereas image and video data are suited for deep learning frameworks to classify human activity. But for Soldier-borne applications, inertial measurement units (IMUs) are an attractive sensor modality as they are small wireless sensors that measure kinematics and can be worn relatively unobtrusive to whoever is using them. However, even using unobtrusive sensors, predicting predict Soldier movements often has unique challenges different than that of the traditional use for HAR to predict Activities of Daily Living (ADL) from ubiquitous sensors like cell phones and smart watches. In HAR in general, inter and intra-person variation in movement can create challenges due to subsequent variation in data. This is especially true for Soldiers who are often asked to perform atypical movements, such as crawling and wall climbing. As such, there are very few open data sets available for this type of analysis and it is unknown how different classification methods will perform with it.

Thus, the goal of this project is to develop and evaluate models for predicting what kind of physical task the person is doing from wearable sensors. First, in this work, we will compare three classifiers method: Random Forest (RF) Classifier , Support Vector Machines (SVM), Gradient Boosted Classifier (GBC) with different features to select the best classifiers method based on accuracy of each classifier.

Background

HAR is a domain specific application of pattern recognition which draws from many years of research and development. As such, a generalized framework, the “Activity Recognition Chain” (ARC), has emerged that outlines the process for taking input data and outputting activity labels.

An ARC comprises stages for data acquisition, signal preprocessing and segmentation, feature extraction and selection, training, and classification. Raw signals are first processed and split into m segments (Wi)

```
from which feature vectors (Xi) are extracted. Given
features (Xi), a model with parameters scores c
activity classes Yi = {y1, ... , yc}...
-Bulling, Blanke, and Schiele 2014
```

Within each stage of the ARC, there are many choices and options for tailoring the process toward a specific task or to accommodate some kind of data. For signal pre-processing, there are many different types of filters and signal reduction techniques. The segmentation stage necessitates choosing window size, step sizes, and percent of window overlaps, all of which can have an impact on model results (Baños et al. 2014) . For feature extraction, there are time domain features, frequency domain features, and even other handcrafted features. Exploratory data analysis of specific data sets, previous research in the specific domain, and domain expertise will craft most of the choices made within the ARC for classification.

The types of activities classified and the types of sensors used through scientific literature is vast (Gupta et al. 2022) . Wearable sensors, smart phones, images and videos, and more sophisticated laboratory equipment like optical motion trackers have been used to classify human activity in a wide variety of environments. For more mobile applications based on wearable sensors, an IMU is a sensor module containing a three-axis accelerometer (linear acceleration) and a three-axis gyroscope (angular velocity) and provide an opportunity to exploit HAR in the real world. Decision tree based classification methods offer an attractive approach to HAR from wearable sensor data (Xu et al. 2017) . Decision trees can often manage with lower computational resources and higher computational speeds due to their rule-based binary decisions. Decision trees can be evaluated at O(log n) time for n-attributes. Gradient boosted classifiers work on decision trees by building an additive model in a forward stage-wise fashion, utilizing the gradient of an arbitrary loss function to optimize the tree parameters. Support vector machines offer a geometry based approach that maximize the margins between distinct classes based on feature distances.

Related Work

RF classification, an ensemble method with many decision trees, is a popular, robust approach that tends to produce high accuracy for HAR. It should be noted that RF classifiers do not make assumptions about distributions and are thus particularly robust to outliers and distribution shifts. In 2020, Nurwulan and Selamaj compared the performance of multiple algorithms on classifying ADL (walking, sitting, standing, etc) from a single smart phone tri-axial accelerometer. The authors even broke down the feature sets from the accelerometer signal to compare time-domain, frequency-domain, and a combination for each classifier. Their RF classifier did outperform other classifiers, with SVM coming in second and K-nearest neighbors (KNN), and multi-layer perceptron following; however, there was only a 1-2% difference in accuracy. In addition, combining frequency-domain features with the time-domain features did not meaningfully affect the performance of any of the classifiers. It should be noted too from this study that the KNN performed well. However, their data set consisted of far fewer features than the current study. Due to the expected high dimensionality of the current project's data set, K-Nearest Neighbors was not attempted. More similar to the current study, Mandha, Devi, and Row (2017) used more (three) IMUs on their subjects as they performed more complex movements - weight lifting tasks. The authors found both a RF and a KNN model to produce 99%+ accuracy.

Many other authors have found similar success using SVMs as their classification algorithm. Sunkad and Soujanya (2016) used the University of Southern California (USC) for Human Activity Recognition database, which includes more dynamic movements like jumping, to assess hyperparameter optimization for SVM. The authors also extracted many time-domain features from a single IMU placed on the pelvis. Kernel shape, regularization parameter, and gamma assessed via cross-validation grid search. Using a radial basis function kernel, with regularization 100, and gamma of 0.001 the authors achieved 99% performance on their data set. In 2022, Kusuma, Minarno, and Safitri took the analysis further by separating the USC HAR database into static (standing, sitting) and dynamic (walking, jumping) movements to determine if SVM performance is affected by what activities are trained on and how the hyperparameters may change. The authors found that splitting the data this way actually improved the accuracy of the SVM on the dynamic set (99%), even beyond the performance on the whole data set (98%).

Thus, the precedent is set that RF and SVM algorithms can attain good performance from low dimensionality data sets with mildly dynamic physical movements. Gradient Boosted classification will also be assessed as a possible alternative to speed up performance on the previous mentioned decision-tree based algorithm, random forest. Each application of HAR is different and the algorithms and parameters optimized for one task may not be optimal for other tasks. As another point, the classification methods described above may require hand crafting features from wearable sensor data which may be difficult for non-subject matter experts. Deep learning methods utilizing convolutional neural networks are common choices for high dimensional data sets such as with multi-sensor IMUs (Wang et al. 2019) . These methods don't require elaborate feature extractions. However, they don't account for the fact that sensor data is time series data and the sequence of samples is important. Recurrent networks, and their variant, the long-short term memory networks, do take data sequence into account and provide a powerful way to model time series data. This paper will not delve into deep learning techniques but they will be explored in later phases of the project.

Project Description

This project will investigate the efficacy of different supervised, multi-class classification models for HAR on a unique data set of Soldier physical tasks. To assess the performance of these classifiers for our application, this project will follow a similar workflow as the ARC described in Bulling, Blanke, and Schiele (2014). Our workflow is depicted in Figure 1.

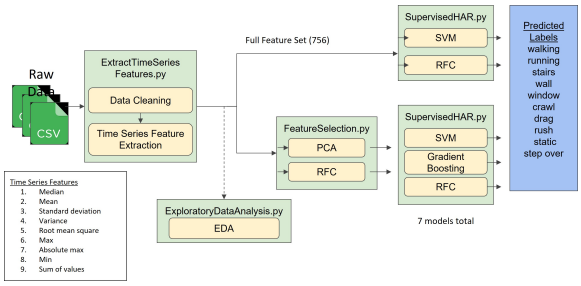

Figure 1: Data pipeline for Soldier movement HAR.

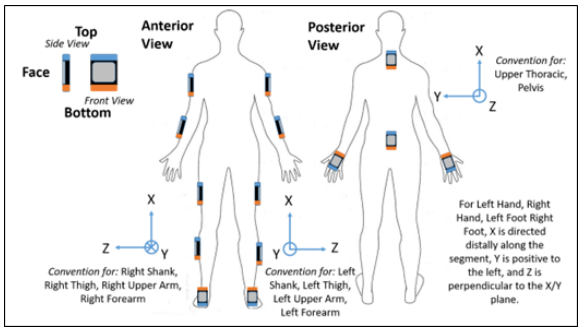

Figure 2: IMU suite for biomechanical analysis during movement.

Data and Computational Resources

All computation was performed in Python 3.10 using a Google Colab Linux x86_64 computing environment (12 GB RAM). Python libraries used are numpy, pandas, matplotlib, scipy, seaborn, random, sklearn, and tsfresh . The data set used consists of IMU data collected from 15 subjects while they completed different 'Soldier' tasks while wearing full body IMU suit. A total of 14 sensors (configuration shown in Fig. 2) were used on different limbs for a total of biomechanical 84 features extracted. The subjects performed 10 different tasks ranging in complexity, walking, running, stairs, static, wall climb, window climb, crawl, dummy drag, bounding rush, and step over. In total, 1160 trials were collected. 142 trials contained missing values. These were either due to malfunctioning sensors or other signal anomalies and as such were removed from analysis resulting in a data set of 1017 trials to learn on. For this specific project, the time series data is stored in tabular form locally in comma separated value (CSV) files. The motion capture software used to collect the data automatically resolves the raw IMU data into different biomechanical features; thus all features denoted similar names like "Right Elbow Flexion".

Feature Extraction and Exploratory Data Analysis

First, the distribution of activity labels displayed a mildly imbalance with the 'walk' trial occurring the most frequently, as depicted in Figure 2. All analysis going forward required a stratified train-test split as well as a stratified k-fold cross validation approach.

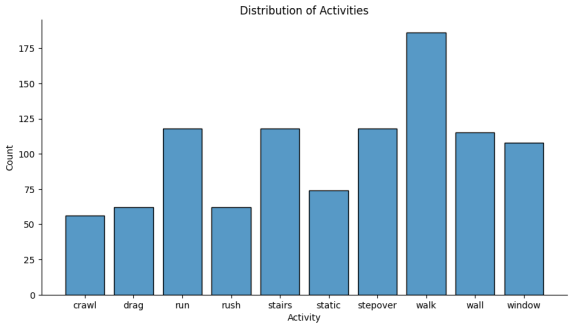

Figure 3: Class distribution across data set.

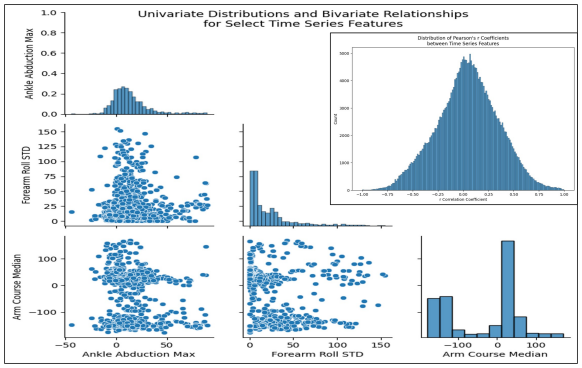

Figure 4: Univariate and Bivariate Distributions and Pearson's R Coefficients.

Ten time-domain statistics were extracted from each time domain variable using the tsfresh package (Christ et al. 2018). These statistics represent aggregate information about the signal over each trial, including signal median, mean, standard deviation, variance, root mean square, max, absolute max, min, signal length, and the sum of values. The signal length variable was removed from further analysis as the trial length is a function of timing of data collection and is not a characteristic of the signal itself. A total of 756 time domain features were extracted and used for further analysis. These feature variables were explored using various visualization techniques to assess distributions and relationships to the target label outcome variable. Insights gained were used to determine best approaches for feature reduction and classification.

Univariate distributions and bivariate relationships within the data set were visualized as histograms and scatter plots. Three features are depicted in Figure 3 as example variables with differing distributions and relationships. As these time domain features are themselves statistics, they can take on non-normal distributions, such as t-student distributions (Ankle Abduction Max, upper right Fig. 3), exponential (Forearm Roll Std, middle column Fig. 3), and multi-modal (Arm Course Median, lower left Fig. 3).

This poses a challenge when trying to implement parametric or linear machine learning algorithms. Indeed, no feature passed a Shapiro-Wilks Test for Normality. Addition-

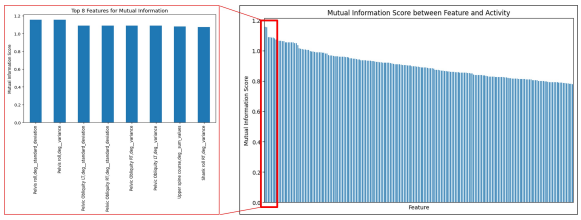

Figure 5: Mutual Information score between each feature and the target activity variable. Call out on the left shows the top 8 features according to mutual information.

ally, through the scatter plots in the lower left triangle of Figure 3, we see that these features often do not have linear relationships. This is further confirmed by the histogram inlet in Figure 3 (upper right hand corner), which shows a significant mass of the inter-feature Pearson's R correlation coefficients fall within -0.25 and 0.25.

Although visualizing the inter-feature relationship is important, for classification tasks understanding the relationship between each feature and the target variable is more important. Mutual information is a metric for quantifying the amount of reduction in uncertainty in a target variable given a known value of another variable or feature. Scores close to 0 indicate little to know mutual information while score 0.7+ indicate the opposite. The mutual information score between each feature and the target activity label is depicted in Figure 4, with a specific call out to the top 8 scoring features. While mutual information can be a good metric for feature selection, most of the features in the data set reach at least 0.8. A more definitive approach to feature reduction is needed.

Feature Reduction

As a note, prior to any feature reduction or further analysis, the data was split with class label stratification into 70% training and 30% testing sets. Because of the high dimensionality of the data set and the varied distributions of the individual features, feature or dimensionality reductions are necessary to make any computational solution tractable. To that end, two feature reduction techniques were applied and assessed (one supervised and one unsupervised approach).

Random Forest A second data set was produced by a supervised approach based on RF feature importance. By fitting a RF model with ground truth labels, feature importance is calculated for each variable as the average decrease in Gini Impurity when that feature is used to split a node. The default threshold of 0.001 was set and any feature with importance less than the threshold was removed from further. The top 10 features with their importance are depicted in Figure 5.

Principal Components Analysis A third data set was produced by an unsupervised approach based on principal components analysis. Based on the cumulative explained variance percentage (Fig. 6), a threshold of 0.9 was set which corresponded to the first 54 principal components. The full feature set was thus transformed into 54 principal components for further performance analysis.

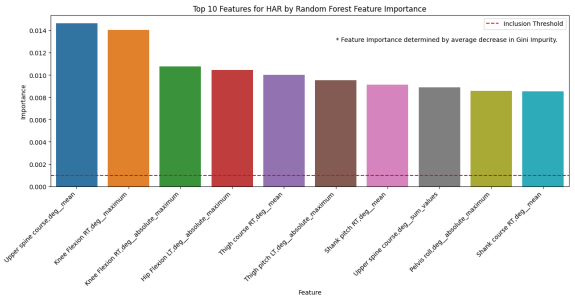

Figure 6: Top 10 features based on Random Forest feature importance scores.

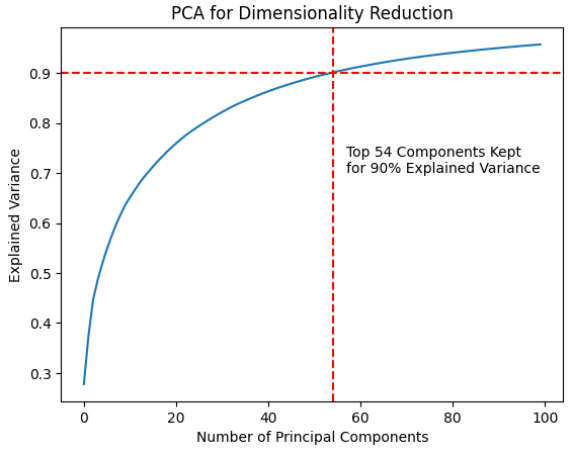

Figure 7: Cumulative sum of explained variance graph for determining number of components for the principal components analysis.

Classification Modeling and Performance Evaluation

The three different models were trained on a different combination of the three different data sets. A total of seven different models were eventually trained and tested (Table 1). The three models were a RF classifier, SVM, and GBC. Each of the models was trained similarly using a 5-fold cross validation technique for hyperparameter tuning. A grid search technique was used to fit and test each combination of hyperparameter (Table 2) for each base model. The set of hyperparameters that produced the highest accuracy for the validation test set was carried forward to compare against the other classification models. Each best tuned model was then used to predict activity labels of the held-out test set. The mean fit time for each tuned model was also compared. Final test set F1 scores and confusion matrices were used to evaluate performance.

Empirical Results

As depicted in the training loss graph (Fig. 6), the best hyperparameters for each base model and feature set scored perfectly when fitting the training data. Subsequent performance on the validation set dropped only slightly across all

	Full	RF Selection	PCA
RF Classifier	•		•
SVM	•	•	•
GBC		•	•

Table 1: Matrix for feature sets and algorithms used in the evaluation.

Model	Hyperparameters
RF Classifier	n_estimators:100, 200, 300; max_features:auto,sqrt,log2; max_depth:6, 8, 10, 12; criterion:gini, entropy
SVM	C:0.1, 1, 10, 100; gamma:1, 0.1, 0.01; kernel:poly, rbf
GBC	learning_rate:0.01, 0.025, 0.05; max_depth:8, 10, 12; n_estimators:100, 200

Table 2: Parameter sets for hyperparameter tuning for each base model.

Figure 8: Accuracy for Training and Validation data sets for all models and feature sets.

models. The model performance data for the test set along with fit times are in Table 4. Best performing model for each metric are bolded. The RF classifier using the full feature set produced the best results for both the validation sets as well as the test set, while only taking less than two seconds longer than the fastest fit time by SVM. Further detail into the performance of each model can be seen in the confusion matrices provided below.

Conclusions and Discussion

The RF classification model with the full feature set produced the highest test F1 score with only minimal increases in training time beyond SVM methods. The SVM model still performs at above 0.93 F1 which is considered very good. These results fall in line with the results from previous IMU-based HAR experiments outlined in "Related Work". Indeed, for our data set, most classifier performed better in part due to the higher dimensionality of the data set and the more distinct movement patterns collected. Gradient Boosting performed the worst and required extensive training time making it intractable for real-world applications.

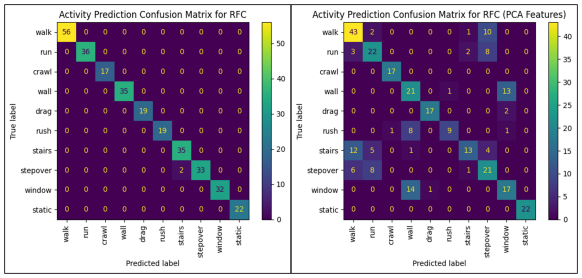

Figure 9: Confusion matrix for Random Forest Classifier performance on test data set.

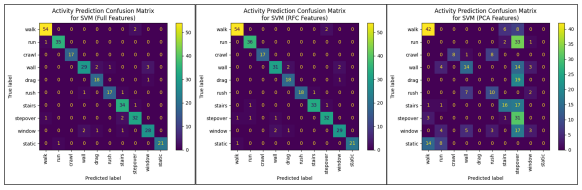

Figure 10: Confusion matrix for Support Vector Machine performance on test data set.

For feature reduction techniques, using the feature importance for down selection did not show any appreciable improvement in performance beyond just using the entire feature set. In contrast, the PCA dimensionality reduction technique produced the worst generalization ability, often reducing the test F1 metric by one half. By only including only the first 54 components covering only 90% of the explain variance, we may be losing important separability in the higher dimensions. Future project can investigate the effect of including more principal components to introduce more variability into the model for better discrimination. Futhermore, deep learning techniques where features are learned from the data and not hand-crafted may offer a more viable solution and should be explored.

Overall, using IMU sensors and time-domain biomechanical features, we can achieve excellent performance when classifying unique Soldier physical tasks. However, the models developed are not suited for real-time classifica-

Model	Feature Set	Mean Fit Time	Precision	Recall	F1
RF Classifier	Full	1.97 ± 0.61	**0.98**	**0.99**	**0.99**
	PCA	2.07 ± 0.75	0.68	0.66	0.66
SVM	Full	0.05 ± 0.01	0.91	0.87	0.93
	RF Selection	$\mathbf{0.05 \pm 0.007}$	0.90	0.90	0.89
	PCA	0.05 ± 0.007	0.71	0.63	0.62
GBC	RF Selection	104.13 ± 4.14	0.89	0.93	0.92
	PCA	30.76 ± 1.84	0.46	0.50	0.45

Table 3: Model performance comparisons. The best performing model for each metric is bolded.

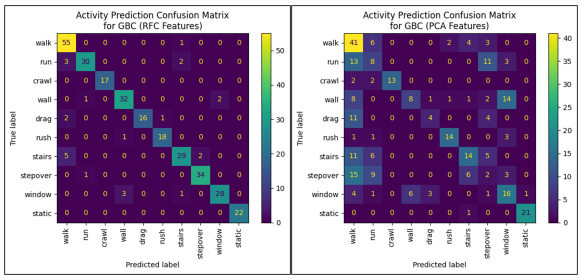

Figure 11: Confusion matrix for Gradient Boosted Classifier performance on test data set.

tion of physical activity. For any practicality in the field, real time methods will need to be developed based on smaller windows of data and faster computation times.

Code Repository

The repository can be found at https://github.com/tweedell/DS5500_Project1 which contains the jupyter notebook where all code and data needed for training and analysis resides.

References

Baños, O.; Gálvez, J. M.; Damas, M.; Pomares, H.; and Rojas, I. 2014. Window Size Impact in Human Activity Recognition. *Sensors (Basel, Switzerland)*, 14: 6474 – 6499.

Bao, L.; and Intille, S. S. 2004. Activity Recognition from User-Annotated Acceleration Data. In Ferscha, A.; and Mattern, F., eds., *Pervasive Computing*, 1–17. Berlin, Heidelberg: Springer Berlin Heidelberg.

Bulling, A.; Blanke, U.; and Schiele, B. 2014. A Tutorial on Human Activity Recognition Using Body-Worn Inertial Sensors. *ACM Comput. Surv.*, 46(3).

Christ, M.; Braun, N.; Neuffer, J.; and Kempa-Liehr, A. W. 2018. Time Series FeatuRe Extraction on Basis of Scalable Hypothesis Tests (Tsfresh – A Python Package). *Neurocomput.*, 307: 72–77.

Gupta, N.; Gupta, S.; and Pathak, R. e. a. 2022. Human Activity Recognition in Artificial Intelligence Framework: a Narrative Review. *Artif Intell Rev*, 55.

Kusuma, W. A.; Minarno, A. E.; and Safitri, N. D. N. 2022. Human activity recognition utilizing SVM algorithm with gridsearch. *AIP Conference Proceedings*, 2453.

Ladha, C.; Hammerla, N. Y.; Olivier, P.; and Plötz, T. 2013. ClimbAX: skill assessment for climbing enthusiasts. *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*.

Mandha, P.; Devi, G. L.; and Row, S. V. 2017. A Random Forest based Classification Model for Human Activity Recognition. *International Conference on Innovative Applications in Engineering and Information Technology(*, 3.

Nurwulan, N. R.; and Selamaj, G. 2020. Random Forest for Human Daily Activity Recognition. *Journal of Physics: Conference Series*, 1655(1).

Tapia, E. M.; Intille, S. S.; and Larson, K. 2004. Activity Recognition in the Home Using Simple and Ubiquitous Sensors. In Ferscha, A.; and Mattern, F., eds., *Pervasive Computing*, 158–175. Berlin, Heidelberg: Springer Berlin Heidelberg.

Wang, J.; Chen, Y.; Hao, S.; Peng, X.; and Hu, L. 2019. Deep learning for Sensor-Based Activity Recognition: A Survey. *Pattern Recognition Letters*, 119.

Xu, L.; Yang, W.; Cao, Y.; and Li, Q. 2017. Human activity recognition based on random forests. In *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 548–553.