Journal of the Midwest Association for Information Systems (JMWAIS)

Volume 2021 | Issue 1 Article 4

2021

Human Activity Recognition: A Comparison of Machine Learning Approaches

LOKNATH SAI AMBATI
loknathsai.ambati@trojans.dsu.edu

Omar El-Gayar
Dakota State University, omar.el-gayar@dsu.edu

Journal of the Midwest Association for Information Systems

Volume 2021 | Issue 1 Article 4

Date: 01-31-2021

Human Activity Recognition: A Comparison of Machine Learning Approaches

Loknath Sai Ambati
Dakota State University, LoknathSai.Ambati@trojans.dsu.edu

Omar El-Gayar
Dakota State University, Omar.El-Gayar@dsu.edu

Abstract

This study aims to investigate the performance of Machine Learning (ML) techniques used in Human Activity Recognition (HAR). Techniques considered are Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, Logistic Regression, Stochastic Gradient Descent, Decision Tree, Decision Tree with entropy, Random Forest, Gradient Boosting Decision Tree, and NGBoost algorithm. Following the activity recognition chain model for preprocessing, segmentation, feature extraction, and classification of human activities, we evaluate these ML techniques against classification performance metrics such as accuracy, precision, recall, F1 score, support, and run time on multiple HAR datasets. The findings highlight the importance to tailor the selection of ML technique based on the specific HAR requirements and the characteristics of the associated HAR dataset. Overall, this research helps in understanding the merits and shortcomings of ML techniques and guides the applicability of different ML techniques to various HAR datasets.

Keywords: Human Activity Recognition, Machine Learning, Performance, Healthcare, Benchmark.

Please note: A prior version of this article received a best paper award submitted for the 2020 Midwest Association for Information Systems (MWAIS) in Des Moines, Iowa which was cancelled due to COVID-19 pandemic. The article has been expanded and received a second round of reviews. We congratulate the authors.

DOI: 10.17705/3jmwa.000065

1. Introduction

The popularity of wearable technology has increased over the recent years (Iqbal et al. 2018). Applications such as self-management aimed at managing disease condition, and self-care for facilitating health and wellbeing have adopted wearable technology to improve health and wellbeing for users. Most of these wearable devices contains sensors such as accelerometers, gyroscopes, magnetometers, heart rate sensors and similar sensors embedded for successful human activity recognition (HAR). The availability of data coupled with the wide ranging applications of HAR resulted in HAR garnering significant attention in academia and in practice (Qin et al. 2020).

In that regard, machine learning and data mining techniques have proved beneficial in extracting features and classifying HAR data (Ramasamy Ramamurthy and Roy 2018). Most of the HAR applications in the market today are striving to improve their performance by utilizing ML techniques and have demonstrated success in terms of performance metrics such as classification accuracy and processing speed (Meyer et al. 2016). Further, HAR data are often characterized by a number of attributes, such as activity type, sensor type, preprocessing steps, and position of sensor on a specific body area. Such diverse characteristics makes HAR particularly challenging and is a persistent driver for ongoing research. Specifically, prior research has mainly focused on developing and improving novel ML models for a number of activities in unique environments and populations (Wang et al. 2019), e.g., elderly individuals in a home care environment (Chen et al. 2017). Further, most of HAR literature is concentrated around improving the HAR performance by considering a single dataset and a specific ML classification technique which limits the generalizability of the findings (Baldominos et al. 2019; Nabian 2017). Although some attempts have been made to compare various ML techniques on multiple HAR datasets (Dohnálek et al. 2014; Li et al. 2018), their focus is often limited to either improving feature learning or finding optimal techniques with the best tradeoff between speed and accuracy rather than a comprehensive approach that could be employed to understand the performance of ML techniques and map them to the characteristics of various HAR data sets.

Accordingly, this research study aims to analyze the performance of different ML classification techniques using various HAR datasets. As HAR sensor data sets can vary significantly with respect characteristics such as sampling frequency, type of activities performed, number of sensors, sensor types and sensor positions, these variations in characteristics have been demonstrated to impact ML techniques hyper parameter tuning, classification performance, and run time. (Baldominos et al. 2019; Dohnálek et al. 2014; Nabian 2017; Wang et al. 2019). This research extends the understanding of the performance of ML classification techniques on HAR data. The significance of this research is both theoretical and practical. From a theoretical point of view, this research helps to understand the merits and shortcomings of ML techniques that could help future researchers figure out how to improve the ML classification techniques for HAR datasets. From a practical point of view, this research helps in guiding the applicability of different ML classification techniques to HAR datasets. Altogether, the research on HAR performance improvement can remarkably facilitate self-management and self-care interventions. In addition, these improvements extend beyond the medical and healthcare domains to other context, wherever the detection of human activity is vital.

The remainder of the paper is organized as follows: a brief literature review is presented in section 2, while section 3 describes the methodology including the characteristics of the dataset and the details of the analysis process. Section 4 illustrates the results obtained from the analysis and section 5 summarizes and discusses the results by comparing with extant literature. Finally, section 6 concludes by summarizing the key contributions, limitations, and suggests directions for future research.

2. Literature Review

2.1 Human Activity Recognition

Raw data obtained from the wearable sensors undergo a number of steps as demonstrated by the Activity Recognition Chain (ARC) model (Bulling et al. 2014) for classifying human activities as shown in Figure 1. In this model, the first step involves sampling the raw data obtained from different sensors with multiple dimensions, before it undergoes preprocessing, segmentation, feature extraction, and finally, classification. Among these steps, feature extraction requires deep domain expertise in the field. Therefore, researchers tend to depend on domain experts for feature engineering and extraction. Utilizing the resultant engineered features with ML and deep learning techniques, the activities are classified into specific human activities (Saha et al. 2018).

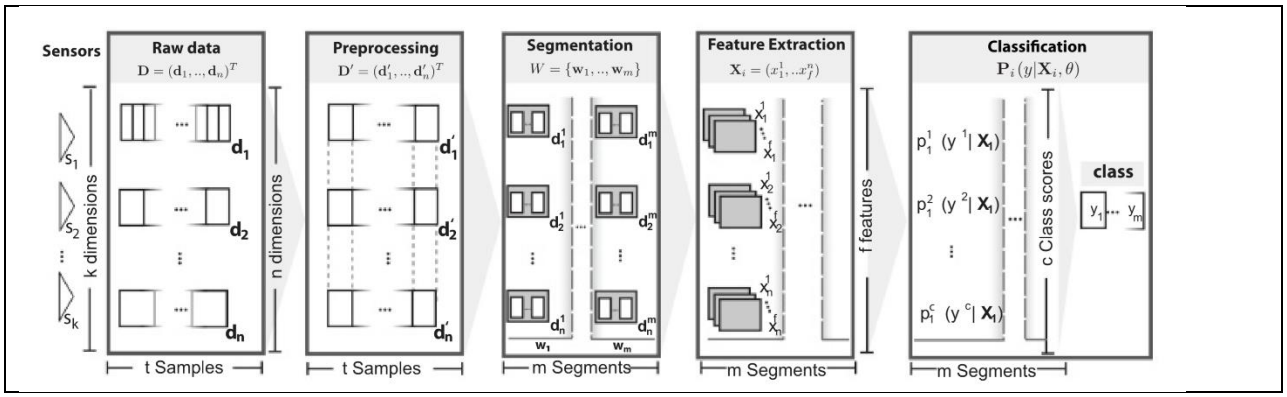

Figure 1: Activity Recognition Chain (ARC) Model (Bulling et al. 2014)

HAR research focuses predominantly on classifying human activities using ML techniques or/and preprocessing the data (Baldominos et al. 2019; Jain and Kanhangad 2018; Nabian 2017; Ronao and Cho 2017; Sousa et al. 2017). HAR using smartphones is a popular sub-field where there is abundant of literature that deals with improving HAR classification using various innovative pre-processing and ML techniques (Anguita et al. 2013; Jain and Kanhangad 2018; Micucci et al. 2017; Nakano and Chakraborty 2017; Ronao and Cho 2014, 2017). Few studies tried to improve the HAR classification obtained from inertial sensors using hyper parameter tuning of ML techniques (Gaikwad et al. 2019; Garcia-Ceja and Brena 2015; Seto et al. 2015). Others have also focused on the problems and difficulties associated to segmentation and proposed solutions to tackle these problems (Kozina et al. 2011; Oresti Banos 12:19:30 UTC). These studies have shown a partial effect of segmentation on the performance of ML techniques. Similarly, few studies demonstrated that the window size affects the HAR classification, e.g., 1-2 second interval results in optimal tradeoff between accuracy and recognition speed (Banos et al. 2014; Ni et al. 2016).

2.2 Comparative analysis

Most of the HAR literature is concentrated around improving the performance HAR using a single dataset and a specific ML technique which limits the generalizability of the findings. There are some studies that tried to consider various ML techniques (Akhavian and Behzadan 2016). These types of analyses compare various ML techniques to identify the most suitable technique for a HAR dataset (Baldominos et al. 2019; Nabian 2017). These studies used a single dataset to understand the relation between few HAR characteristics such as sensor position, type of activity and hyper parameter tuning on ML performance. Recent studies on HAR has shown evidence that no prior preprocessing of raw sensor data has shown reasonable ML performance especially in a comparative study (Dohnálek et al. 2014). Although some attempts have been made to compare various ML techniques on multiple HAR datasets (Dohnálek et al. 2014; Li et al. 2018), their focus is often limited to either improving feature learning methods or finding optimal techniques with the best tradeoff between speed and accuracy. Accordingly, there is a need for a comprehensive study to evaluate and benchmark the performance of various ML techniques with different HAR datasets and map the characteristics of various HAR datasets to appropriate ML techniques. Prior work on HAR data partly tried to address this gap by comparing multiple HAR datasets with the accuracy score of different ML techniques (Ambati and El-Gayar 2020). We aim to extend this work by collectively considering multiple HAR datasets, the type of activities being classified, the performance of an expanded portfolio of ML techniques, and the use of an expanded set of performance metrics to get more insights in understanding the ML techniques and their relation to HAR data in conjunction with the extant literature. These insights can help future researchers in designing a robust and comprehensive framework/model depending on the HAR application.

3. Methodology

3.1 Datasets

We used three HAR datasets in a manner that captures the diversity of characteristics commonly present in various datasets. The first two datasets (Pampa2 and mHealth) are from the University of California, Irvine (UCI) data repository.

The datasets were chosen in such a way that they are distinct in terms of sensors utilized, sampling frequency, activity environment and similar attributes, and are utilized in prior research (Anguita et al. 2013; Gaikwad et al. 2019; Garcia-Ceja and Brena 2015; Nakano and Chakraborty 2017). This makes these datasets unique and appropriate for utilizing them to benchmark various ML techniques. The third dataset is selected from the SWELL project supported by the Dutch national program COMMIT (Shoaib et al. 2014). Table 1 presents the data sets and their characteristics. 3D accelerometers, 3D gyroscope, and 3D magnetometer are the common sensors employed in all the three datasets. These sensors have become a basic functionality for wearable devices that attempt to recognize human activity. 3D accelerometer helps in recognizing the speed with which the user is moving in all three dimensions, 3D magnetometer helps in recognizing the orientation of the user with respect to earth's magnetic north, and 3D gyroscope helps in recognizing the angular velocity of the user. Other than these three sensors, each dataset has some unique sensors when compared to each other. For example, Pamap2 dataset has a heart rate monitor and a temperature sensor, while mHealth has an ECG sensor and SWELL has a linear acceleration sensor. With respect to data collection, Pamap2 relies on wireless IMU's, while mHealth uses wearables, and SWELL uses smartphones to collect the data. All activities in a particular dataset are conducted for approximately the same amount of time and are represented evenly in the data sets. Therefore, data imbalance does not constitute an issue. When data size is considered, Pampap2 dataset is the largest dataset with a 519,185-record size, while SWELL stands second with a 189,000-record size, and MHealth being the smallest with 102,959 record size.

Dataset	Sensors	Sensor Position	Activities performed	Dataset description	Sampling Frequency
Pamap2	3 Colibri wireless IMUs (inertial measurement units) and BM-CS5SR (HR monitor) – Accelerometer, Gyroscope, magnetic sensor and temperature sensor.	wrist, chest and side ankle.	lying, sitting, standing, walking, running, cycling, Nordic walking, watching TV, computer work, car driving, ascending stairs, descending stairs, vacuum cleaning, ironing, folding laundry, house cleaning, playing soccer and rope jumping.	9 subjects (1 female and 8 male) aged 27.22 (+-) 3.31 years performed the 12 mandatory activities and 6 optional activities for 2-3 minutes.	100 samples/sec
Mhealth	accelerometer, a gyroscope, a magnetometer and ECG (Shimmer2 [BUR10] wearable sensors).	chest, right wrist and left ankle	L1: Standing still (1 min), L2: Sitting and relaxing (1 min), L3: Lying down (1 min), L4: Walking (1 min), L5: Climbing stairs (1 min), L6: Waist bends forward (20x), L7: Frontal elevation of arms (20x), L8: Knees bending (crouching) (20x), L9: Cycling (1 min), L10: Jogging (1 min), L11: Running (1 min), L12: Jump front & back (20x)	10 volunteers of diverse profile performed 12 physical activities for about 1 min	50 samples/sec
SWELL	accelerometer, a gyroscope, a magnetometer, and a linear acceleration sensor (Samsung Galaxy SII (i9100) smartphone).	upper arm, wrist, two pockets, and belt position	walking, sitting, standing, jogging, biking, walking upstairs and walking downstairs	10 participants performed 7 activities for 3-4 minutes. All are male with ages 25-30.	50 samples/sec

Table 1. Dataset Characteristics

3.2 Analysis

We compared different ML techniques using a number of HAR datasets (Pamap2, mHealth and SWELL) across various ML performance metrics. The ML techniques are Naïve Bayes, Support Vector Machine (SVM) with linear kernel, K-Nearest Neighbor (KNN), Logistic Regression, Stochastic Gradient Descent (SGD), Decision Tree, Decision Tree with entropy, Random Forest, Gradient Boosting Decision Tree (XGBoost), and NGBoost algorithm. Although, deep learning techniques such as neural network based algorithms are attracting popularity over the recent years, they tend to over fit in the case of HAR data (Jobanputra et al. 2019). Moreover, the runtime of each dataset over various ML techniques is already high when run on Python Jupyter Notebook using eight-generation intel i7 processor considering the data is not extensively preprocessed, therefore, applying neural network-based techniques on these large HAR datasets would significantly increase runtime. Considering these circumstances, neural network techniques are not implemented in this research. Although accuracy is the most popular ML performance metric in HAR (Li et al. 2018), we utilized additional metrics such as precision, recall, F1 score, support and runtime to facilitate an in-depth analysis. Table 2 shows the description of ML metrics employed for evaluating the ML performance.

All the three datasets are minimally preprocessed by addressing the missing values and excluding the data during the transient stage (transition from one activity to another) based on timestamp and standardizing the format of the data such

that it would be easier to implement the ML techniques and interpret the results obtained from the analysis. Specifically, the data is standardized in a manner such that each row represents the sample values for each sensor for a specific sampling time, and each column represents a sensor except for timestamp, subject ID, and classification activity. After standardizing the format of each dataset with minimal preprocessing, we split the data into training (70%) and test data (30%) for each dataset. Once all the datasets are split into training and test data, we train (including hyperparameters tuning) of the various ML techniques using each of the datasets. Then, we evaluate each ML technique with the considered performance metrics on each dataset.

ML Metrics	Accuracy	Precision	Recall	F1 score	Predicting Run Time
Definition	The ratio of number of correct predictions to the total number of predictions.	The ratio of number of correctly predicted positive values to the total predicted positive values.	The ratio of correctly predicted positive values to the total number of positive values.	Harmonic mean of precision and recall.	Time taken for target classification using test data.
Formula	(TP+TN)/(TP+TN+FP+FN)	TP/(TP+FP)	TP/(TP+FN)	2*(Precision*Recall)/ (Precision+Recall)	-

Table 2. Description of ML metrics

Where:

- True Positive (TP) - Number of correctly predicted positive values.
- True Negative (TN) - Number of correctly predicted negative values.
- False Positive (FP) – Number of predictions that interpret negative values as positive values.
- False Negative (FN) – Number of predictions that interpret positive values as negatives.

To investigate the potential tradeoff between classification performance and prediction runtime, we identify the Pareto efficient ML techniques for each of the datasets. Pareto efficiency is a concept where no individual criterion can be declared better without a sacrifice in one of the other criterion (Bokrantz and Fredriksson 2017). Accuracy is used as the metric representing classification performance, while prediction run time is measured in seconds.

3.2.1 Classification performance by activity

To get a better understanding of the performance of the various ML techniques using the various datasets, we considered three cases as follows.

Individual activities: In this case, the target variable is represented as a categorical variable where each category representing one activity such as sitting, standing, running, and lying for each dataset. This type of grouping is very popular in comparative analysis for understanding each activity and the respective effect of ML technique. It provides for the most fidelity as all activities are accounted for. However, it allows for the number of activities (classes) to vary among the three datasets which may compound the comparative analysis of the performance of various ML techniques.

Grouped activities: To address the aforementioned issue and considering that differentiating between sitting and standing, and between walking fast and running, and similar differentiation can be very difficult to obtain (Gjoreski et al. 2014), we conducted another set of experiments where all the activities in each dataset are divided into two categories namely locomotive activities and stationary activities. Activities where the user is staying idle with no physical movement such as sitting, standing, and lying are considered stationary activities. All other activities which require the user to perform a physical movement such as, but not limited to walking, running, jumping, climbing stairs and similar activities are categorized as locomotive activities. It is assumed that since all the locomotive activities share a similarity that the sensor movement is dynamic and similarly, stationary activities share a similarity that all the sensor movement would be idle, it should alleviate the problem of differentiating similar activities. This categorization should guide us towards understanding more towards the type of activity and how it is going to affect the ML techniques and their performance, respectively.

Common activities: Another possibility for standardizing the activities across datasets while maintaining as much fidelity as possible (in terms of the number of activities/classes) considered, we conducted an additional experiment where we included the activities that are common in all three datasets. The common activities in all the datasets are walking, sitting, standing, running, cycling, and climbing stairs.

4. Results

4.1 Classification of individual activities

Table 3 depicts the performance of the various ML techniques on the three data sets. With respect to accuracy, the performance of ML techniques irrespective of the datasets in the order of best performance are XGboost, Random Forest, KNN, SVM, Decision Tree with entropy, Decision Tree, NGBoost, Logistic Regression, Naïve Bayes, and SGD. There are three exceptions to this observation, Naïve Bayes technique performed better in the case of SWELL when compared to Logistic Regression, SGD performed better in the case of Mhealth when compared to Naïve Bayes technique, and Random Forest, KNN, and SVM performed better than XGboost in the case of mHealth.

With respect to precision, recall, and F1 Score, the performance of ML techniques irrespective of the datasets in the order of best performance are KNN, SVM, Random Forest, XGboost, Decision Tree with entropy, Decision Tree, Logistic Regression, Naïve Bayes, and SGD. There are two exceptions to this observation, Naïve Bayes technique performed better than Logistic Regression in the case of SWELL and SGD performed better than Naïve Bayes in the case of the mHealth dataset.

The performance of ML techniques irrespective of the datasets in the order of least runtime, Logistic Regression, SGD, Decision Tree with entropy, Decision Tree, Random Forest, Naïve Bayes, XGboost, SVM, and KNN. There is one exception to this observation, XGboost has lower run-time when compared to Naïve Bayes in the case of Pamap2 dataset.

		Naïve Bayes	SVM	KNN	SGD	Logistic Reg.	DT	DT with Entropy	RF	XGBoost	NGBoost
Accuracy	Pamap2	0.901	0.999	0.999	0.9	0.92	0.999	**0.999**	0.999	0.999	0.936
	SWELL	0.879	0.996	0.998	0.847	0.855	0.975	0.977	0.995	0.999	0.881
	MHealth	0.521	0.965	0.991	0.629	0.738	0.911	0.918	0.939	0.934	0.875
Precision	Pamap2	0.91	1	1	0.9	0.92	1	1	1	1	0.93
	SWELL	0.88	1	1	0.85	0.85	0.98	0.98	1	1	0.88
	MHealth	0.52	0.97	0.99	0.63	0.72	0.91	0.92	0.94	0.94	0.87
Recall	Pamap2	0.90	1	1	0.9	0.92	1	1	1	1	0.93
	SWELL	0.88	1	1	0.85	0.86	0.98	0.98	1	1	0.87
	MHealth	0.52	0.97	0.99	0.63	0.74	0.91	0.92	0.94	0.93	0.87
F1 Score	Pamap2	0.90	1	1	0.9	0.92	1	1	1	1	0.93
	SWELL	0.88	1	1	0.85	0.85	0.98	0.98	1	1	0.88
	MHealth	0.55	0.97	0.99	0.62	0.65	0.91	0.92	0.94	0.93	0.87
Predicting Run-Time (s)	Pamap2	7.281	639.2	12,860	0.145	0.15	0.182	**0.149**	1.518	6.923	401.765
	SWELL	1.026	303.271	6,527	0.062	0.046	0.087	0.072	0.599	2.638	281.942
	MHealth	0.398	232.583	488.51	0.021	0.02	0.03	0.024	0.262	2.625	122.888

Table 3. Performance metrics of ML techniques for individual activities

4.2 Classification of grouped activities

As shown in Table 4, all ML techniques performs better using the Pamap2 dataset on all performance metrics except the runtime when compared to the other two datasets. When we consider accuracy, precision, recall and F1 score, the general trend that is followed by the ML techniques irrespective of the datasets is, Logistic Regression, SGD, Naïve Bayes, NGBoost Decision Tree, Decision Tree with entropy, SVM, KNN, Random Forest, and XGboost. There is one exception where Logistic Regression performed better than Naïve Bayes in the case of Pampap2 dataset.

When prediction run-time is considered, the general trend followed by the ML techniques irrespective of the dataset in the order of the shortest to longest runtime is Logistic Regression, SGD, Decision Tree with entropy, Decision Tree, Random Forest, Naïve Bayes, XGboost, SVM, NGBoost, and KNN. As expected, (with only two classes), the performance metrics of ML techniques for grouped activities is better when compared to the individual activities.

		Naïve Bayes	SVM	KNN	SGD	Logistic Reg.	DT	DT with Entropy	RF	XGBoost	NGBoost
Accuracy	Pamap2	0.966	1	0.999	0.999	0.999	1	1	1	1	1
	SWELL	0.99	0.999	0.999	0.97	0.97	0.998	0.998	0.999	0.999	0.998
	MHealth	0.989	0.991	0.999	0.834	0.814	0.998	0.998	0.999	0.999	0.996
Precision	Pamap2	0.97	1	1	1	1	1	1	1	1	1
	SWELL	0.99	1	1	0.97	0.97	1	1	1	1	0.99
	MHealth	0.99	0.99	1	0.83	0.80	1	1	1	1	0.99
Recall	Pamap2	0.97	1	1	1	1	1	1	1	1	1
	SWELL	0.99	1	1	0.97	0.97	1	1	1	1	0.99
	MHealth	0.99	0.99	1	0.83	0.81	1	1	1	1	0.99
F1 Score	Pamap2	0.97	1	1	1	1	1	1	1	1	1
	SWELL	0.99	1	1	0.97	0.97	1	1	1	1	0.99
	MHealth	0.99	0.99	1	0.82	0.80	1	1	1	1	0.99
Predicting Run-Time (s)	Pamap2	0.712	21.831	12965	0.09	0.053	0.102	0.108	0.596	1.058	118.688
	SWELL	0.343	11.216	6505.16	0.025	0.029	0.048	0.044	0.265	0.57	46.751
	MHealth	0.069	52.799	486.397	0.006	0.008	0.014	0.012	0.109	0.217	10.065

Table 4. Performance metrics of ML techniques for grouped activities

4.3 Classification of common activities

As shown in Table 5, all ML techniques perform better using the Pamap2 dataset on all performance metrics except the runtime when compared to the other two datasets. When we consider accuracy, precision, recall and F1 score, the general trend that is followed by the ML techniques irrespective of the datasets is, SGD, Logistic Regression, Naïve Bayes, NGBoost Decision Tree, Decision Tree with entropy, SVM, KNN, Random Forest, and XGboost. There is one exception where Logistic Regression performed better than Naïve Bayes in the case of Pampap2 dataset.

When prediction run-time is considered, the general trend followed by the ML techniques irrespective of the dataset in the order of the shortest-longest runtime is Logistic Regression, SGD, Decision Tree with entropy, Decision Tree, Random Forest, Naïve Bayes, XGboost, SVM, NGBoost, and KNN. The performance metrics of ML techniques for common activities are better when compared to the individual activities and slightly lower when compared with the grouped activities.

		Naïve Bayes	SVM	KNN	SGD	Logistic Reg.	DT	DT with Entropy	RF	XGBoost	NGBoost
Accuracy	Pamap2	0.945	0.999	0.999	0.984	0.987	0.999	0.999	1	1	0.999
	SWELL	0.939	0.998	0.999	0.924	0.931	0.991	0.993	0.998	0.999	0.964
	MHealth	0.931	0.99	0.998	0.789	0.829	0.988	0.99	0.999	0.999	0.959
Precision	Pamap2	0.95	1	1	0.98	0.98	1	1	1	1	0.99
	SWELL	0.94	1	1	0.92	0.93	0.99	0.99	1	1	0.96
	MHealth	0.93	0.99	1	0.78	0.82	0.99	0.99	1	1	0.95
Recall	Pamap2	0.95	1	1	0.98	0.99	1	1	1	1	0.99
	SWELL	0.94	1	1	0.92	0.93	0.99	0.99	1	1	0.96
	MHealth	0.93	0.99	1	0.79	0.83	0.99	0.99	1	1	0.95
F1 Score	Pamap2	0.95	1	1	0.98	0.98	1	1	1	1	0.99
	SWELL	0.94	1	1	0.92	0.93	0.99	0.99	1	1	0.96
	MHealth	0.93	0.99	1	0.78	0.82	0.99	0.99	1	1	0.95
Predicting Run-Time (s)	Pamap2	1.374	50.455	3603.943	0.076	0.074	0.082	0.078	0.661	1.987	305.698
	SWELL	0.959	120.28	4645.316	0.05	0.046	0.077	0.066	0.514	2.221	203.605
	MHealth	0.121	31.922	152.443	0.01	0.008	0.013	0.013	0.109	0.557	28.638

Table 5. Performance metrics of ML techniques for common activities

5. Discussion

When Pamap2 dataset is employed, all the ML techniques performed to their best for all possible groupings of activities. The relatively large size of the data resulting from the higher sampling frequency, and the additional sensors (temperature sensor and heart rate monitor) utilized in the dataset, positively affected ML performance. This leads us to conclude that the overall improvement of ML performance metrics tends to be associated with the number of sensors and higher sampling rate employed to collect the HAR data. Generally, tree-based algorithms such as Random Forest, NGBoost, and XGboost outperformed Naïve Bayes, SGD and Logistic Regression (with an exception of KNN and SVM, as their runtime is very high for real time usage) in terms of ML performance metrics thereby attesting to the claims made by other studies that Tree based techniques perform better than other techniques in the field of HAR (Sánchez and Skeie 2018).

When performance metrics of ML techniques for individual activities and common activities are compared with grouped activities, all the ML techniques performed better when activities are grouped as locomotive or stationary activities. This supports the assertion that it is particularly challenging to differentiate similar activities among a particular group (stationary and locomotive). Although, this observation is expected, this comparison provides an additional dimension for comparing ML techniques behavior.

Although, previous studies achieved accuracies up to 0.97 and F1 score of 0.84 with just the wrist position using deep learning techniques (Baldominos et al. 2019), this study obtained much higher accuracies and F1 scores using less complex ML techniques compared to neural network based techniques. However, in every dataset utilized, a combination of three or more sensor positions were employed. Therefore, evaluating the performance for each sensor position separately would give more insights but drastically increases the complexity of the analysis given the additional consideration for the number and location of sensors. This can be further explored in the future research.

When run-time is analyzed, it is expected that the dataset having more data (both in terms of features as well as data collected) would take more time to run a particular ML technique. Accordingly, in all considered scenarios, the predicting run time for any ML technique is highest when Pamap2 dataset is employed, followed by SWELL, and mHealth. Usually, all the ML techniques take more time for training and takes less time for predicting. Naïve Bayes technique on the other hand, took the least time for training the model but took a relatively long time for prediction using the test data. Naïve Bayes model size (with respect to the number of model parameters to be estimated) is relatively small compared to the other ML techniques considered. Moreover, depending on the conditional independence assumption being true, the model converges very fast resulting in a low training run time and less data being required for training compared to the other datasets considered (Mark 2015). If we consider any real time application that requires recognition of human activity, the main concern for designing the application would be minimizing prediction run time while maintaining acceptable classification performance. This puts Naïve Bayes ML technique at a disadvantage. Further, Logistic Regression tend to have the shortest run time while KNN has the longest run time in most of the cases considered.

Considering the tradeoff between classification performance represented using accuracy we find that DT with Entropy appears on the pareto efficient frontier regardless of the data set. DT with entropy is also the sole ML technique that is Pareto efficient for the Pamap2 dataset. Random Forest and Logistic Regression are Pareto efficient for SWELL and MHealth, while XGBoost is Pareto efficient for SWELL only. The prevalence of tree-based ML techniques such as DT with entropy, Random Forest, and to some extent XG Boost further supports prior research tree based techniques perform better than other techniques in the field of HAR (Sánchez and Skeie 2018). Although, KNN has a relatively long run time, it exhibits the best classification performance irrespective of the HAR dataset. Interestingly, KNN and SVM are in effect Pareto efficient for MHealth. However, run time for these two techniques is in the order of four magnitude larger than the run time for the other techniques rendering a steep tradeoff between run time and classification performance. This, considering very high run time of SVM and KNN, neither of these techniques may be suitable for real time applications.

If an application is more interested in the accuracy, low on budget for additional sensors such as heart rate monitor, then XGBoost would be an ideal solution with some run time tradeoff compared to Random Forest. Similarly, if an application is more interested in short run time, then Decision Tree with entropy would be an ideal solution with a small tradeoff with the accuracy. But in most of the other cases, DT with Entropy is the optimal performer for any combination of weighted performance metrics selected. There can always be some exceptions such as in an application with very limited data, then KNN or XGBoost might be a better fit depending on the size of data.

In essence, depending on the requirements of the HAR application and data amount, we can choose the sensor types (Shoaib et al. 2014), sensor positions (Baldominos et al. 2019), ML techniques (Dohnálek et al. 2014), sampling frequency (Wang et al. 2019), and similar characteristics based on the insights provided in this research. The key insights pointed out in this study:

- It is relatively difficult to differentiate similar activities among a particular group (stationary and locomotive).
- High sampling frequency improves the ML performance metrics (Accuracy, precision, recall, F1 score, and support), however, it will take a toll on the run-time.
- DT with Entropy stands out to be the optimal performer in most cases of the HAR applications.
- Ensemble techniques outperforms traditional ML techniques in terms of ML performance metrics except run time for HAR data.
- Naïve Bayes technique is efficient when there are more activities involved.
- Naive Bayes technique takes the least time for training the data and build the model but takes a heavy toll in time taken for predicting the test data.

6. Conclusion

In this study, the performance of various ML techniques used for HAR are evaluated using ML performance metrics such as accuracy, precision, recall, F1 score, and run time on multiple HAR datasets. We investigated the relationship of different HAR dataset characteristics to the performance of various ML techniques. Examples include the amount of the data collected, sampling frequency, sensor types, type of activity performed, number of activities performed, and sensor positions. Although, DT with Entropy performed best on most types of HAR data considering its performance metrics across all the datasets, there is no single silver bullet for HAR data. The findings highlight the importance to tailor the selection of ML technique based on the specific HAR requirements and the characteristics of the associated HAR dataset. Future research can analyze the impact of sensor types and positions individually on ML performance. Another potential future research avenue of this study is extending the portfolio of ML techniques to include an investigation of deep learning and (more importantly in the context of wearables, light-weight architectures). Future research could also explore the effect of various pre-processing to further explore the Pareto efficient frontier between run-time performance and classification performance.

7. References

Akhavian, R., and Behzadan, A. H. 2016. "Smartphone-Based Construction Workers' Activity Recognition and Classification," *Automation in Construction* (71), pp. 198–209. (https://doi.org/10.1016/j.autcon.2016.08.015).

Ambati, L. S., and El-Gayar, O. 2020. "A Comparative Study of Machine Learning Approaches for Human Activity Recognition," in *Proceedings of the Fifteenth Midwest Association for Information Systems Conference*, Des Moines, Iowa, May 28, p. 6.

Anguita, D., Ghio, A., Oneto, L., Parra, X., and Reyes-Ortiz, J. L. 2013. "A Public Domain Dataset for Human Activity Recognition Using Smartphones," in *ESANN*.

Baldominos, A., Cervantes, A., Saez, Y., and Isasi, P. 2019. "A Comparison of Machine Learning and Deep Learning Techniques for Activity Recognition Using Mobile Devices," *Sensors* (19:3), p. 521. (https://doi.org/10.3390/s19030521).

Banos, O., Galvez, J.-M., Damas, M., Pomares, H., and Rojas, I. 2014. "Window Size Impact in Human Activity Recognition," *Sensors (Basel, Switzerland)* (14:4), pp. 6474–6499. (https://doi.org/10.3390/s140406474).

Bokrantz, R., and Fredriksson, A. 2017. "Necessary and Sufficient Conditions for Pareto Efficiency in Robust

Multiobjective Optimization," *European Journal of Operational Research* (262:2), pp. 682–692. (https://doi.org/10.1016/j.ejor.2017.04.012).

Bulling, A., Blanke, U., and Schiele, B. 2014. "A Tutorial on Human Activity Recognition Using Body-Worn Inertial Sensors," *ACM Comput. Surv.* (46:3), 33:1-33:33. (https://doi.org/10.1145/2499621).

Chen, O. T.-, Tsai, C., Manh, H. H., and Lai, W. 2017. "Activity Recognition Using a Panoramic Camera for Homecare," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, , August, pp. 1–6. (https://doi.org/10.1109/AVSS.2017.8078546).

Dohnálek, P., Gajdoš, P., and Peterek, T. 2014. "Human Activity Recognition: Classifier Performance Evaluation on Multiple Datasets," *Journal of Vibroengineering* (16:3), pp. 1523–1534.

Gaikwad, N. B., Tiwari, V., Keskar, A., and Shivaprakash, N. C. 2019. "Efficient FPGA Implementation of Multilayer Perceptron for Real-Time Human Activity Classification," *IEEE Access* (7), pp. 26696–26706. (https://doi.org/10.1109/ACCESS.2019.2900084).

Garcia-Ceja, E., and Brena, R. 2015. "Building Personalized Activity Recognition Models with Scarce Labeled Data Based on Class Similarities," in *Ubiquitous Computing and Ambient Intelligence. Sensing, Processing, and Using Environmental Information*, Lecture Notes in Computer Science, J. M. García-Chamizo, G. Fortino, and S. F. Ochoa (eds.), Springer International Publishing, pp. 265–276.

Gjoreski, H., Kozina, S., Luštrek, M., and Gams, M. 2014. "Using Multiple Contexts to Distinguish Standing from Sitting with a Single Accelerometer," in *European Conference on Artificial Intelligence (ECAI)*.

Iqbal, Z., Ilyas, R., Shahzad, W., and Inayat, I. 2018. "A Comparative Study of Machine Learning Techniques Used in Non-Clinical Systems for Continuous Healthcare of Independent Livings," in *2018 IEEE Symposium on Computer Applications Industrial Electronics (ISCAIE)*, , April, pp. 406–411. (https://doi.org/10.1109/ISCAIE.2018.8405507).

Jain, A., and Kanhangad, V. 2018. "Human Activity Classification in Smartphones Using Accelerometer and Gyroscope Sensors," *IEEE Sensors Journal* (18:3), pp. 1169–1177. (https://doi.org/10.1109/JSEN.2017.2782492).

Jobanputra, C., Bavishi, J., and Doshi, N. 2019. "Human Activity Recognition: A Survey," *Procedia Computer Science* (155), The 16th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2019),The 14th International Conference on Future Networks and Communications (FNC-2019),The 9th International Conference on Sustainable Energy Information Technology, pp. 698–703. (https://doi.org/10.1016/j.procs.2019.08.100).

Kozina, S., Lustrek, M., and Gams, M. 2011. *Dynamic Signal Segmentation for Activity Recognition*.

Li, F., Shirahama, K., Nisar, M. A., Köping, L., and Grzegorzek, M. 2018. "Comparison of Feature Learning Methods for Human Activity Recognition Using Wearable Sensors," *Sensors* (18:2), p. 679. (https://doi.org/10.3390/s18020679).

Mark. 2015. "How to Decide When to Use Naive Bayes for Classification," *Data Science, Analytics and Big Data Discussions*, , October 31. (https://discuss.analyticsvidhya.com/t/how-to-decide-when-to-use-naive-bayes-for-classification/5720, accessed July 13, 2020).

Meyer, J., Schnauber, J., Heuten, W., Wienbergen, H., Hambrecht, R., Appelrath, H., and Boll, S. 2016. "Exploring Longitudinal Use of Activity Trackers," in *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, , October, pp. 198–206. (https://doi.org/10.1109/ICHI.2016.29).

Micucci, D., Mobilio, M., and Napoletano, P. 2017. "UniMiB SHAR: A Dataset for Human Activity Recognition Using Acceleration Data from Smartphones," *Applied Sciences* (7:10), p. 1101.

(https://doi.org/10.3390/app7101101).

Nabian, M. 2017. "A Comparative Study on Machine Learning Classification Models for Activity Recognition," *Journal of Information Technology & Software Engineering* (7:4), pp. 4–8. (https://doi.org/10.4172/2165-7866.1000209).

Nakano, K., and Chakraborty, B. 2017. "Effect of Dynamic Feature for Human Activity Recognition Using Smartphone Sensors," in *2017 IEEE 8th International Conference on Awareness Science and Technology (ICAST)*, , November, pp. 539–543. (https://doi.org/10.1109/ICAwST.2017.8256516).

Ni, Q., Patterson, T., Cleland, I., and Nugent, C. 2016. "Dynamic Detection of Window Starting Positions and Its Implementation within an Activity Recognition Framework," *Journal of Biomedical Informatics* (62), pp. 171–180. (https://doi.org/10.1016/j.jbi.2016.07.005).

Oresti Banos. 12:19:30 UTC. *Evaluating the Effects of Signal Segmentation on Activity Recognition*, Science. (https://www.slideshare.net/orestibl/banos-iwbbio-2014pdf).

Qin, Zhen, Zhang, Y., Meng, S., Qin, Zhiguang, and Choo, K.-K. R. 2020. "Imaging and Fusing Time Series for Wearable Sensor-Based Human Activity Recognition," *Information Fusion* (53), pp. 80–87. (https://doi.org/10.1016/j.inffus.2019.06.014).

Ramasamy Ramamurthy, S., and Roy, N. 2018. "Recent Trends in Machine Learning for Human Activity Recognition—A Survey," *WIREs Data Mining and Knowledge Discovery* (8:4), John Wiley & Sons, Ltd, p. e1254. (https://doi.org/10.1002/widm.1254).

Ronao, C. A., and Cho, S. 2014. "Human Activity Recognition Using Smartphone Sensors with Two-Stage Continuous Hidden Markov Models," in *2014 10th International Conference on Natural Computation (ICNC)*, , August, pp. 681–686. (https://doi.org/10.1109/ICNC.2014.6975918).

Ronao, C. A., and Cho, S.-B. 2017. "Recognizing Human Activities from Smartphone Sensors Using Hierarchical Continuous Hidden Markov Models," *International Journal of Distributed Sensor Networks* (13:1), p. 1550147716683687. (https://doi.org/10.1177/1550147716683687).

Saha, S. S., Rahman, S., Rasna, M. J., Zahid, T. B., Islam, A. K. M. M., and Ahad, M. A. R. 2018. "Feature Extraction, Performance Analysis and System Design Using the DU Mobility Dataset," *IEEE Access* (6), pp. 44776–44786. (https://doi.org/10.1109/ACCESS.2018.2865093).

Sánchez, V. G., and Skeie, N.-O. 2018. "Decision Trees for Human Activity Recognition Modelling in Smart House Environments," *SNE Simulation Notes Europe* (28:4), pp. 177–184. (https://doi.org/10.11128/sne.28.tn.10447).

Seto, S., Zhang, W., and Zhou, Y. 2015. "Multivariate Time Series Classification Using Dynamic Time Warping Template Selection for Human Activity Recognition," in *2015 IEEE Symposium Series on Computational Intelligence*, , December, pp. 1399–1406. (https://doi.org/10.1109/SSCI.2015.199).

Shoaib, M., Bosch, S., Incel, O. D., Scholten, H., and Havinga, P. J. M. 2014. "Fusion of Smartphone Motion Sensors for Physical Activity Recognition," *Sensors* (14:6), Multidisciplinary Digital Publishing Institute, pp. 10146–10176. (https://doi.org/10.3390/s140610146).

Sousa, W., Souto, E., Rodrigres, J., Sadarc, P., Jalali, R., and El-Khatib, K. 2017. "A Comparative Analysis of the Impact of Features on Human Activity Recognition with Smartphone Sensors," in *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web*, WebMedia '17, New York, NY, USA: ACM, pp. 397–404. (https://doi.org/10.1145/3126858.3126859).

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), pp. 3–11. (https://doi.org/10.1016/j.patrec.2018.02.010).

Author Biographies

Loknath Sai Ambati is a doctoral student pursuing Ph.D degree in Information Systems with specialization in analytics and decision support at Dakota State University. Loknath is currently working as graduate research assistant at Dakota State University, his research interests include Health Information Technology, Population Health, Health Informatics, and Data Analytics. His work was published in IGI Global, IIS, AMCIS and MWAIS. Besides, he is an active referee of several international journals and conferences like Journal of Intelligent and Fuzzy Systems, IACIS, AMCIS, and HICSS.

Omar El-Gayar, Ph.D. is a Professor of Information Systems at Dakota State University. Dr. El-Gayar has an extensive administrative experience at the college and university levels as the Dean for the College of Information Technology, United Arab Emirates University (UAEU) and the Founding Dean of Graduate Studies and Research, Dakota State University. His research interests include: analytics, business intelligence, and decision support with applications in problem domain areas such as healthcare, environmental management, and security planning and management. His inter-disciplinary educational background and training is in information technology, computer science, economics, and operations research. Dr. El-Gayar's industry experience includes working as an analyst, modeler, and programmer. His numerous publications appear in various information technology related fields. Dr. El-Gayar serves as a peer and program evaluator for accrediting agencies such as the Higher Learning Commission and ABET, as a panelist for the National Science Foundation, and as a peer-reviewer for numerous journals and conferences. He is a member of a number of professional organizations such as the Association for Information Systems (AIS) and the Association for Computing Machinery (ACM).