EAI Endorsed Transactions
on Context-Aware Systems and Applications
Research Article EAI.EU

Gradient Descent Machine Learning with Equivalency Testing for Non-Subject Dependent Applications in Human Activity Recognition

T.A. Woolman[1,*], J.L. Pickard[2]

[1]Indiana State University, Terre Haute, Indiana, USA
[2]East Carolina University, Greenville, North Carolina, USA

Abstract

INTRODUCTION: A solution to subject-independent HAR prediction through machine learning classification algorithms using statistical equivalency for comparative analysis between independent groups with non-subject training dependencies.
OBJECTIVES: To indicate that the multinomial predictive classification model that was trained and optimized on the one-subject control group is at least partially extensible to multiple independent experiment groups for at least one activity class.
METHODS: Gradient boosted machine multinomial classification algorithm is trained on a single individual with the classifier trained on all activity classes as a multinomial classification problem.
RESULTS: Levene-Wellek-Welch (LWW) Statistic calculated as 0.021, with a Critical Value for LWW of 0.026, using an alpha of 0.05.
CONCLUSION: Confirmed falsifiability that incorporates reproducible methods into the quasi-experiment design applied to the field of machine learning for human activity recognition.

Keywords: Human activity recognition, digital sensor, telemetry, gradient boosting, gradient descent, machine learning, classification, statistical equivalence testing

Received on 18 April 2022, accepted on 13 July 2022, published on 15 July 2022

doi: 10.4108/eetcasa.v8i24.1996

*Corresponding author. Email: twoolman@ontargettek.com

1. Introduction

Increasingly sophisticated network sensor telemetries, coupled with the accumulating 4G and 5G network (4th and 5th generation mobile network standards) coverage across major population centers are providing unique opportunities for the classification of complex human activities utilizing an ensemble suite of wireless sensors intended for generic, cross-functional applications. Smart devices such as smartphones and smartwatches are now ubiquitous in our daily lives. The combination of pervasive high-speed and high-bandwidth wireless network infrastructures and robust, reliable fast-cycle commoditized sensor suites embedded into commonplace smartphone and smart watch devices present opportunities for human activity recognition (HAR) research from network devices that only a few years ago would not have been attainable (1).

The multitude of sensors (accelerometer, gyroscope, magnetometer, and GPS unit) in smart devices work together to accurately calculate a person's movement and position, making them useful for HAR (2), (3), and (4). Using sensor data obtained from smart swatches and smart phones to identify and classify various activities is the focus of many HAR studies in the extant literature. For example, Zhang, et al. (5) used smartphones successfully for activity recognition of construction workers, (6) used smartphones to estimate indoor localization, and (7) used

wearable sensors to for detection of the activities of a tennis player.

Once data collection has taken place, the raw data undergoes pre-processing to remove any unwanted data, known as noise, and to segment the data into "windows" of time duration. Next, Machine Learning (ML) and Deep Learning (DL) models are used to classify the collected data. Recent research utilizing ML and DL has been largely successful in interpreting some of these machine-generated telemetry datasets for automated predictive classification tasks, with reasonably high accuracy. However, these approaches are often individually tailored to train models on a specific individual resulting in subject-dependent prediction of HAR. The studies of (8) and (9) note that human activity patterns are dependent on the individual, meaning that people have diverse activity styles which is a technical challenge in the HAR field. For practical HAR use-case scenarios it is sometimes the case that the participants are unknown to the system, especially if such a HAR system is to be extensible to be able to accurately classify activities for a broader population beyond the initial training study cohort.

This study proposes a solution to the challenge of subject-independent prediction of HAR through a novel combination of machine learning classification algorithm ensembles while focusing on the empirical use of a test of statistical equivalency for the comparative analysis of classification model results between independent groups with non-subject training dependencies (distinct time-series observations across multiple human test subjects for multiple classes of activities). For this study we use a public biometric dataset of smartphone and smartwatch activity developed by the Wireless Sensor Data Mining (WISDM) Lab at Fordham university. The dataset is comprised of six kinds of daily human activities (walking, jogging, upstairs, downstairs, sitting, and standing) and includes 1,098,207 pieces of data (1).

The WISDM dataset was utilized because of the well-curated nature of the dataset that provides a baseline for this research. WISDM consists of multiple sensor types recording data over three axes, spanning multiple activity classes. As such, WISDM represents a significant "state of the art" in curated datasets that are freely available for HAR prediction model research and model comparative performance analysis studies.

Model performance was analyzed across a population of n-subjects for the purposes of concluding with linear and/or nonlinear statistical equivalence that ensemble algorithm models are both falsifiable and externally valid for use across multiple groups with equivalent net positive predictive capability for a stated statistical power, alpha test statistic and equivalency boundary. This will in effect demonstrate a potential for the models to thus be "universal" across groups (human subjects) within at least one net positive predictive activity class for accurate HAR classification equivalency within the given domain of specified HAR activities.

1.1. Statement of the Problem

Previous research on human activity recognition (HAR) using artificial intelligence (AI) and machine learning (ML) has focused primarily on accurately predicting classes of known, human-labeled activities for each observation of network device telemetry data for specific human subjects. These models have typically been trained on labeled activities from specific subjects, and as such generally have subject dependency recognition constraints and their accuracy in identifying activities depends in large part on the feature selection process (1), (10), and (11). These studies have indicated that there are potentially several machine learning and deep learning algorithms that can be applied with varying degrees of success to achieve some degree of reliability for multinomial HAR classification accuracies.

The problem addressed in this study is that much of the current state of the art for HAR classification has focused on data collection and limited time series analysis of HAR classifications, without addressing issues related to falsification and external validity for the AI and ML classification findings for independent groups outside of the subject-dependent training data cohort. While HAR research has been significant in moving the state of the art forward, the rapid growth in the adoption and use of smart phone and smart watch platforms connected directly to unrestrained Internet connections has increased the capability for collecting smart device sensor telemetry data at a very high observation velocity, potentially allowing the introduction of more robust HAR classification methods that are both falsifiable and externally validated across multiple groups outside of the training cohort limitation.

This study is unique for two primary reasons, in that it both addresses a gradient descent methodology that is successfully applied with high accuracy to aspects of the HAR predictive classification problem with the WISDM dataset, and furthermore addresses the application of falsifiability to the research question. The falsifiability application applies a novel application of a test of statistical equivalence to addressing a null hypothesis statement using data obtained from a nonparametric statistical learning algorithm.

1.2. Null Hypothesis and Alternative Hypothesis

In a traditional experiment with hypothesis testing, such as in a Mann-Whitney U-test, when treatment conditions are analyzed between groups that have continuous type dependent variables that are non-normally distributed, the primary goal of the analysis of the null hypothesis is to reject the statement that no difference between the control and experiment groups exists after the effect of some intervention or treatment in favor of the alternative hypothesis. Thus, the alternative hypothesis would ideally

be supported by a rejection of the null hypothesis, in which the research provides evidence that a statistically significant difference between groups of independent observations of the mean (two-tailed test) or median (one-tailed test) of the dependent variable of a continuous data type is present between multiple experiment groups.

However, the goal of this research is to indicate that the multinomial predictive classification model that was trained and optimized on the one-subject control group for the WISDM dataset is at least partially extensible to multiple independent experiment groups for at least one activity class representing a nominal dependent factor variable, with both sensitivity and selectivity within an equivalence boundary for an upper and lower limit for k independent groups. Thus, rather than the use of a standard null and alternative hypothesis, the researchers propose the use of a statistical test of equivalence as discussed by (12), by testing for noninferiority and two-sided equivalence of paired variables in independent, randomly selected groups of unequal membership sizes.

By demonstrating equivalence in this manner for the independent experiment group outcomes, the researchers intend to strengthen the scientific validity of multinomial classification research as it is applied to the use of machine learning algorithms by applying falsifiability and external validity to classification models with reduced subject dependency recognition constraints, greatly expanding the potential application for these classifiers by incorporating tests of statistical equivalence in digital sensor telemetry HAR AI/ML applications.

H0: The classification predictions for the k independent experiment groups for a given activity class with net positive prediction accuracy with non-subject dependency falls outside the equivalence interval.

H1: The classification predictions for the k independent experiment groups for a given activity class with net positive prediction accuracy with non-subject dependency falls are equivalent within the equivalence boundary.

By being able to successfully reject the null hypothesis, the authors hope to demonstrate an increased replicability and external validity to the proposed classification methodology for human activity recognition experiments for independent groups with non-subject dependency, with the primary goal of improving the overall utility and falsifiability for HAR machine learning classification methods.

1.3. Limitations

This post-hoc study relies upon a previously collected and human-labelled curated dataset (WISDM) and as such has no ability to augment the data to provide extensible attributes that could enhance causal inferences for hypothesis testing. Key limitations therefore include the finite sample size of unknown statistical power as well as the time series durations and scope of the curated dataset, and the finite number and type of dependent variable class factors chosen for the WISDM dataset. Other limitations include the assumed reliability of the instrumentation used to collect the longitudinal observational telemetry data from the human subjects in the WISDM dataset, as calibration and validation of sensors was not specifically addressed by Weiss (2019).

2. Methodology

Because of the complex nature of smart device telemetry traffic and a potentially increasing number of threat actor scenarios related to law enforcement activities for suspect behaviors and other first responder interests, a combination of traditional statistical analysis techniques and supervised machine learning multinomial classification algorithms were used to create models for linear and nonlinear feature extraction to perform multinomial classification with probability scoring of the known labeled HAR classes across groups that are of interest to this research. From the models produced, a new framework was established to allow the use of a "universal" set of HAR classification predictions for specific types of smart phone and smart watch sensor telemetries that can be applied across groups within a network traffic universe with a statistically significant probability value against a stated test statistical threshold.

The data collected for this research utilized the WISDM dataset produced by (1), a curated structured dataset containing observations across four classes of device telemetries, including smart phone accelerometers and gyroscopes, and smart watch accelerometers and gyroscope data. Data was recorded in approximate 3-minute windows across the x, y and z axis for all sensors. Likewise, transformations of this sensor data were made for each observation window by Weiss, including distribution binning for each axis, average sensor values for each axis, time between peaks for each of the three axes, the standard deviation for each axis and the variance value for each axis.

Additionally, a resultant value was calculated in this original dataset, consisting of the square of the sum for each resulting x, y and z value per observation and then averaging those values across all observations. Likewise, XY, XZ and YZ cosine distances between sensor values were recorded, along with XY, XZ and YZ correlation values for each observation.

Finally, 13 values per axis per observation were recorded that represented the short-term power spectrum of a wave function. This was calculated by the dataset author based on a linear cosine transformation from a log power spectrum from a non-linear mel scale. The relationship of the resultant independent variable to each of the 18 classes of the activity dependent variable for subject ID “1600” (the initialized model training subject) is shown for each of the four sensor types in Figures 1 through 4.

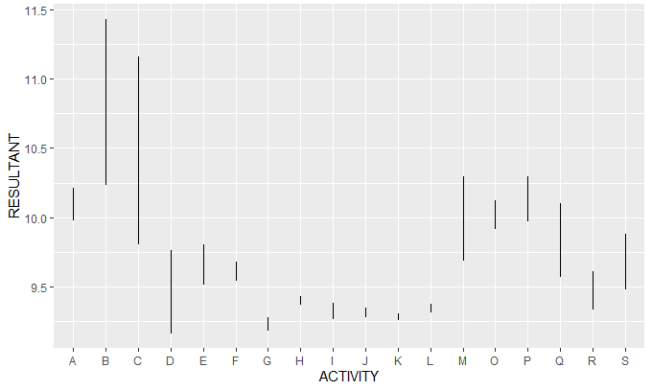

Figure 1. Resultant value by activity class for phone accelerometer, participant ID 1600

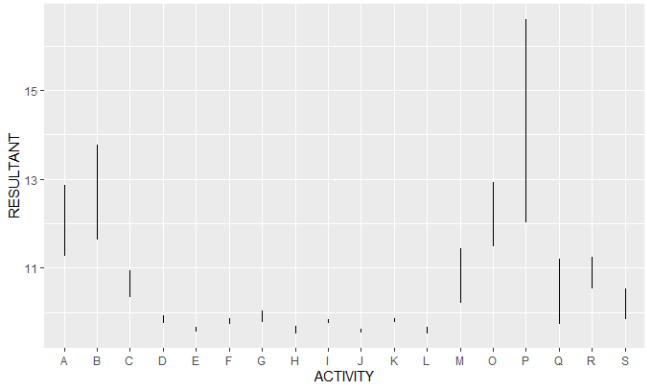

Figure 2. Resultant value by activity class for watch accelerometer, participant ID 1600

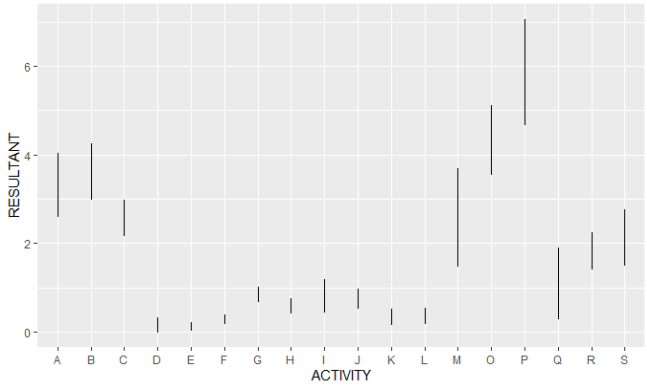

Figure 3. Resultant value by activity class for watch gyroscope, participant ID 1600

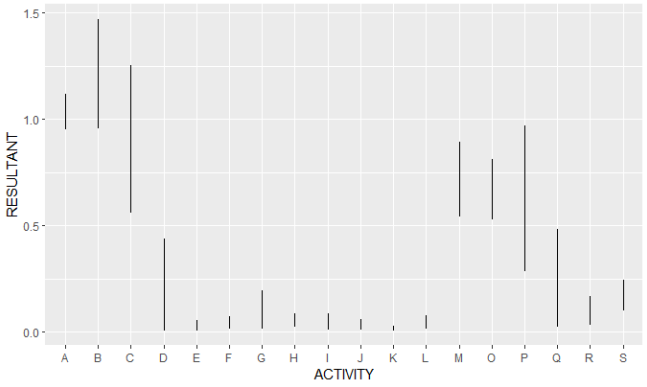

Figure 4. Resultant value by activity class for phone gyroscope, participant ID 1600

The subjects were identified by a unique ID code in the original dataset, and the human- labeled activity class, dependent variable, (DV) was then recorded for each observation. A new, additional independent variable was added for each observation in this study to indicate the sensor class ("phoneaccel", "phonegyro", "watchaccel" or "watchgyro"), for purposes of assisting the machine learning algorithm to further differentiate variations in the telemetry data for each activity class.

Observations for each distinct participant were combined into individual group datasets, consisting of smart watch accelerometer, smart watch gyroscope, smart phone accelerometer and smart watch gyroscope data plus the sensor classification label and the dependent variable (human labeled activity classifier). There were 18 distinct activity classifier labels used as the dependent variable for the multinomial classifier, including walking, jogging, climbing stairs, sitting, standing, typing, brushing teeth, eating soup, eating chips, eating pasta, drinking, eating a sandwich, kicking a ball, catching a ball, dribbling a ball, writing, clapping and lastly, folding clothes (1).

A total of 94 columns (93 independent variables including the newly added nominal, categorical variable for sensor type, plus 1 dependent variable) exists in the dataset for each unique study participant, consisting of a mixture of continuous ratio statistical data types and qualitative, categorical (nominal) data types. The sensor data was collected at a rate of 20Hz or every 50 ms, during the approximate 3 minutes per activity per study participant window.

A gradient boosted machine multinomial classification algorithm was then trained on a single individual from the available data universe, with the classifier trained on all 18 activity classes as a multinomial classification problem. Gradient boosted machines (GBM) are defined by (13) as a machine learning algorithm that incorporates a gradient-descent function as a component of a boosting method, where the procedure consecutively fits new models to provide a more accurate estimate for a response variable. Thus, the principle for this algorithm is primarily the construction of a new set of base-learners that are

maximally correlated to the negative gradient of a loss function for a given ensemble.

This initially trained GBM multinomial classification model was used as a benchmark for attempting to predict 5 additional participants HAR telemetries (k = 5 independent groups) in the study, with each participant having a similar number of recorded telemetry observations from all sensor combinations in each activity class. The number of total observations for each study participant are shown in Table 1.

Table 1. Observation counts for each of the k independent groups used from the WISDM dataset

Subject ID	Distinct Observations from all sensors for all activities
1600 (Training Subject)	1296
1601 (Training Subject)	1460
1602 (Training Subject)	1393
1603 (Training Subject)	1462
1604 (Training Subject)	1290
1605 (Training Subject)	1471
1606 (Training Subject)	1290
1607 (Training Subject)	1424
1608 (Training Subject)	1483
1609 (Training Subject)	1306
1610 (Training Subject)	1610
1611 (Training Subject)	1611

2.1. Experimental Design

An 80%/10%/10% train/test/validation random split was conducted for the initial (baseline) model, using only participant ID 1600 through 1605. This provided 5 distinct, independent groups of observations containing a total of 6901 telemetry activities The goal of this initial baseline model was to attempt to train the GBM classifier with enough information from each sensor class with the various metric attributes and the sensor type IV to learn to recognize activities from a large sample of telemetry events from multiple independent groups. This baseline classification model was then used to attempt to predict activities on other subjects that the model was not trained on in the study to produce a participant-independent predictive classification model.

Such a participant-independent model that could then be applied equally with significant equivalence to the other selected study participants from the available dataset universe. This could then potentially demonstrate falsifiable research by rejecting the null hypothesis and thus support a scientific finding utilizing machine learning classification models for HAR.

The cohort of selected independent groups from the universe of available WISDM data was divided into two major classes, training subjects and prediction subjects. This is not to be confused with a control group and an experiment group, as the null hypothesis statement utilized in this experiment does not seek to measure the effect of a treatment or intervention in a quasi-experiment outcome.

Rather, this experiment rather seeks to measure the ability of the predictive model to be statistically equivalent on predicting k independent cohort groups for a specific activity class, when said k independent groups work have not been utilized for training the supervised machine learning model. Thus, the experiment seeks to utilize HAR activity training data from the distinct training subject cohort and apply it with a test of statistical equivalence with negligible change in effectiveness across human prediction subjects (independent groups) that the model was not exposed to for supervised learning.

A Leven-Wellek-Welch (LWW) statistical test of equivalence was then utilized to determine if a negligible difference existed in the independent experiment groups that the subject-dependent GBM classifier model was trained on, for the purpose of attempting to reject the null hypothesis statement. The LWW test was conducted utilizing methods defined by (14), utilizing the Negligible library package for R (15) and the associated neg.indvars function to conduct a negative effect test for variances of multiple independent group populations. The GBM multiclass prediction algorithm used was from h2o version 3.36.0 as initially developed by (16), for the R statistical programming language, R (17). An α test statistic value of 0.05 was used for this LWW test of equivalence based on its use as a traditional cutoff value for Type I error. A statistical power (1 - β) of 0.975 was achieved for the test of equivalence for k independent groups using the stated α and an Epsilon equivalence margin of 0.25, for a minimum sample size per group of 505 and a minimum total sample size of 3030.

3. Results

A multinomial classification GBM model using the h2o package in R (R Core Team, 2021) was trained on WISDM data with the addition of the sensor categorical independent variable for subject ID 1600 using a multinomial response distribution for ntrees = 6000, seed = 123, max_depth = 4, a learn_rate of 0.001 and min_rows = 3 parameter set.

This trained GBM classifier then produced impressive accuracies on a test and validation split utilizing n-fold

cross validation (k=5) for the subject-dependent control group data. Performance metrics across all 18 activity classes for this baseline classification prediction model were as follows, using metrics commonly used for multiclassification (multinomial) predictive models (MSE, RMSE, Logloss, Mean Per-Class Error and R2):

Mean Error Rates for the Subject-Dependent (Control Group) GBM Model, 10% Training Partition Test

MSE: 0.08864071

RMSE: 0.2977259

Logloss: 0.3531052

Mean Per-Class Error: 0.09345439

R^2: 0.9967421

Mean Error Rates for the Subject-Dependent (Control Group) GBM Model, 10% Validation Partition Test

MSE: 0.1937875

RMSE: 0.440213

Logloss: 0.8683309

Mean Per-Class Error: 0.2081449

R^2: 0.9933996

Activity class "B", which was the jogging activity, was noted as being particularly accurate in this model with 100% accurate class label predictions taking place for the train and validation partition tests for this baseline classification model. Thus, the jogging activity became the focus of this prediction equivalency research for the subject-independent experiment groups.

Applying this subject-dependent trained model to the independent experiment groups that were untrained, the classification error for the jogging activity increased but an overall strong relative accuracy was maintained across all experiment groups for this class. Group ID 1605 produced an error rate of 0.3537, Group ID 1607 produced an error rate of 0.4024, Group ID 1608 produced an error rate of 0.3333, Group ID 1615 produced an error rate of 0.4189 and lastly, Group ID 1624 produced an error rate of 0.6389.

Using a Negligible Effect Test for Variances of Independent Populations test in the R statistical programming environment with an LWW test based on Cribbie, et al., the following results were achieved by analyzing the 5 randomly selected experiment group population variances, using an assumption of independence, using a "conservative" Epsilon value of 0.25, as defined by (1):

Group Variances:

0.2314062, 0.2434508, 0.2244898, 0.2467605, 0.2339593

Group Standard Deviations:

0.481047, 0.4934073, 0.4738035, 0.4967499, 0.4836934

Group Median Absolute Deviations:

0, 0, 0, 0, 0

Ratio of Largest to Smallest Variances:

1.099206

Thus, the Levene-Wellek-Welch (LWW) Statistic calculated as 0.02158431, with a Critical Value for LWW determined as being 0.02651899, using an alpha statistic of 0.05. The Critical Value for LWW is below the stated alpha test statistic. The Null Hypothesis Statistical Test decision, based on the null hypothesis that the differences between the population variances falls outside the equivalence interval, can be rejected. The k independent experiment groups are statistically equivalent within an equivalency boundary as defined by the stated Epsilon parameter and the stated alpha test statistic, for our statistical power in this equivalence experiment.

4. Conclusion

In this research the authors have demonstrated that it is possible to apply tests of falsifiability that incorporate reproducible methods into the quasi-experiment design and apply this to the field of machine learning for human activity recognition. Specifically, the authors successfully applied a novel technique to demonstrate statistical equivalence within a stated equivalence boundary range, for the predictive classification performance of a set of groups that were outside of the subject-dependent training data for the model in question. To further support statistical reliability, the number of independent groups that were tested for statistical equivalence were based on a statistical power test.

While the independent groups that the predictive classification GBM model was trained on for HAR activity classification performed very well in a train/test/split cross-validation test, the ultimate focus of this research was to attempt to broadly apply aspects of this trained HAR classification model and apply it to independent groups of participants that the model was not trained on. In limited HAR activity classes, the model performed adequately at predicting specific HAR activities from complex, multi-sensor device ensembles that provided a very broad range of time series telemetry-based data sets.

Further, the authors based their primary research on the use of gradient boosted machine algorithms, rather than utilizing deep learning methodologies. The authors

speculate that GBM models, when properly optimized, performed well in this use case because the WISDM dataset provided a relatively long duration of telemetry data for each observation cycle, thereby reducing the utility of convolutional and other forms of deep learning neural networks such as LSTM that attempt to recall prior observation attempts in the aggregate time series for each participant.

Based on the results obtained to date with this research, the authors speculate that with a larger dataset for the same activity classes, with more training subjects and many more observations for each training class, further HAR telemetry classification advancements can be made. The authors believe that it is deemed likely that a larger number of HAR classes could be successfully predicted for human subjects that were independent of the training dataset cohort using this same algorithm and research methodology shown. This novel combination of methodology and algorithm application therefore shows significant potential for non-subject dependence HAR telemetry predictions across a large population using tests of statistical equivalency.

5. Recommendations for Further Research

A variety of opportunities for further research present themselves from this study. One key opportunity involves investigating the effects of increased longitudinal study observations across the range of class factors in the dependent variable. Increasing the number of observations for these classes could substantially improve the potential accuracy especially for classes that have similar biomechanical motions, such as typing versus writing, or eating chips versus eating a sandwich.

Likewise, incorporating additional sensors from the smartwatch or smartphone devices such as pedometers, proximity sensors and ambient light sensors could also potentially provide key response indicator variables that may be useful independent variables that provide statistically useful predictive effect for some factors of the dependent variable.

Additional research potential may also exist in utilizing the methodology for nonparametric falsifiability studies utilized in McAlexander and Mentch (2020). In that study, causal inferences for a series of research questions were determined using nonlinear statistical learning models and applied to standard hypothesis testing methods and confidence intervals within a parametric framework. This was accomplished when regularity conditions were produced through subsampling of the nonparametric model predictions.

This allowed for the capture of complex nonlinear relationships in the data between the responses and the predictor that would otherwise have been largely invisible to a traditional parametric regression model, while still maintaining the ability to provide statistically significant findings of cause and effect. This approach may require the development of a larger longitudinal study for the various dependent variable classes to make appropriate use of this technique, based on methods discussed by Zhang and Yuan for minimal sample size and statistical power (2018).

References

1. Weiss GM. Wisdm smartphone and smartwatch activity and biometrics dataset. UCI Machine Learning Repository: WISDM Smartphone and Smartwatch Activity and Biometrics Dataset Data Set. 2019 Sep;7:133190-202.
2. Amezzane I, Fakhri Y, El Aroussi M, Bakhouya M. Towards an efficient implementation of human activity recognition for mobile devices. EAI Endorsed Transactions on Context-aware Systems and Applications. 2018 Mar 14;4(13).
3. Voicu RA, Dobre C, Bajenaru L, Ciobanu RI. Human physical activity recognition using smartphone sensors. Sensors. 2019 Jan;19(3):458.
4. Shoaib M, Bosch S, Incel OD, Scholten H, Havinga PJ. Fusion of smartphone motion sensors for physical activity recognition. Sensors. 2014 Jun;14(6):10146-76.
5. Zhang M, Chen S, Zhao X, Yang Z. Research on construction workers' activity recognition based on smartphone. Sensors. 2018 Aug;18(8):2667.
6. Kang J, Lee J, Eom DS. Smartphone-based traveled distance estimation using individual walking patterns for indoor localization. Sensors. 2018 Sep;18(9):3149.
7. Benages Pardo L, Buldain Perez D, Orrite Uruñuela C. Detection of tennis activities with wearable sensors. Sensors. 2019 Jan;19(22):5004.
8. San Buenaventura CV, Tiglao NM, Atienza RO. Deep Learning for Smartphone-Based Human Activity Recognition Using Multi-sensor Fusion. InInternational Wireless Internet Conference 2018 Oct 15 (pp. 65-75). Springer, Cham.
9. Chen K, Zhang D, Yao L, Guo B, Yu Z, Liu Y. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. ACM Computing Surveys (CSUR). 2021 May 22;54(4):1-40.
10. Irfan S, Anjum N, Masood N, Khattak AS, Ramzan N. A Novel Hybrid Deep Learning Model for Human Activity Recognition Based on Transitional Activities. Sensors. 2021 Jan;21(24):8227.
11. Solorio-Fernández S, Carrasco-Ochoa JA, Martínez-Trinidad JF. A survey on feature selection methods for mixed data. Artificial Intelligence Review. 2021 Sep 29:1-26.
12. Wellek S. Testing statistical hypotheses of equivalence. Chapman and Hall/CRC; 2002 Nov 12.
13. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Frontiers in neurorobotics. 2013 Dec 4;7:21.
14. Jabbari Y, Cribbie R. Negligible interaction test for continuous predictors. Journal of Applied Statistics. 2021 Feb 20:1-5.
15. Cribbie, R., Udi, A., Beribiski, N., Chalmers, P., Counsell, A., Farmus, L., Gutierrez, N., Ng, V. (2022). Negligible: A Collection of Functions for Negligible Effect / Equivalence Testing. CRAN - Package negligible (r-project.org). Accessed 2 Mar 2022.
16. Aiello, S., Kraljevic, T., & Maj, P. (2015). Package 'h2o'. *dim*, *2*, 12.

17. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
18. McAlexander, R. J., & Mentch, L. (2020). Predictive inference with random forests: A new perspective on classical analyses. *Research & Politics*, *7*(1), 2053168020905487.
19. Zhang, Z., & Yuan, K.-H. (2018). *Practical Statistical Power Analysis Using Webpower and R* (Eds). Granger, IN: ISDSA Press.