ARL-TR-9347 ● Nov 2021

Time-Series Classification for Predicting Self-Reported Job Performance

by Alexander F Danvers, Matthias R Mehl, Esther M Sternberg, and Evan C Carter

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-9347 ● Nov 2021

Time-Series Classification for Predicting Self-Reported Job Performance

Alexander F Danvers and Evan C Carter
Human Research and Engineering Directorate,
DEVCOM Army Research Laboratory

Matthias R Mehl and Esther M Sternberg
University of Arizona

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
November 2021	Technical Report	1 July 2020–29 September 2021

4. TITLE AND SUBTITLE
Time-Series Classification for Predicting Self-Reported Job Performance

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Alexander F Danvers, Matthias R Mehl, Esther M Sternberg, and Evan C Carter

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
DEVCOM Army Research Laboratory
ATTN: FCDD-RLH-FA
Aberdeen Proving Ground, MD 21005

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-9347

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES
ORCID IDs: Alexander F Danvers, 0000-0002-7307-9384; Matthias R Mehl, 0000-0003-2698-5007; Esther M Sternberg, 0000-0001-5867-3556; Evan C Carter, 0000-0001-7471-8769

14. ABSTRACT
Widely available wearable devices allow for large-scale passive sensing of human behavior in real-world contexts. Previous research has established that features of heart rate are correlated with constructs related to workplace performance, such as stress and task engagement. In this study, we leverage a large-scale longitudinal study (N = 212, days = 60) to develop machine learning models to predict self-reported job performance. Using the Random Convolutional Kernel Transform (ROCKET) algorithm, which approximates convolutional layers in neural networks with low computational cost, estimates of daily heart rate, heart rate variability, physical activity, social activity, and day of the week were used to predict four job performance outcome measures. Results indicate the ROCKET algorithm had a large proportional increase in the accuracy of baseline models (48% to 121% increase), but the overall levels of accuracy for the best models remained modest (Matthew's Correlation Coefficient ~ 0.10).

15. SUBJECT TERMS
job performance, machine learning, heart rate variability, mobile sensing, daily life

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Evan C Carter
Unclassified	Unclassified	Unclassified	UU	28	19b. TELEPHONE NUMBER (Include area code) (240) 478-9295

Contents

List of Figures

List of Tables

1. Introduction

The purpose of the work described here is to develop algorithms that can accurately predict daily job performance using widely available mobile sensing technology. Because data on physiology and physical activity can be easily collected from a variety of off-the-shelf devices, large-scale deployment of passive sensing technologies is feasible for Soldiers, as well as for large sectors of the civilian workforce. Such data holds the potential for conversion into detailed, daily insights into job performance, which could be used to assess and potentially diagnose problems in performance as they arise. Here, we develop machine learning algorithms for predicting job performance from the dynamics of daily behavior.

The identification of correlates of job performance has a long history in industrial/organizational psychology, the logic being that such models would facilitate interventions or decision-making to improve outcomes. Early efforts employed survey measurements made at a single point in time; for example, a meta-analysis of situational judgment tests found they correlated at reasonably high levels with job performance metrics in cross-sectional samples (McDaniel et al. 2001). Cognitive ability, personality factors like conscientiousness and emotional stability, and emotional intelligence have all also been related to job performance (Hurtz and Donovan 2000; Salgado 2003; Tracey et al. 2007; Joseph et al. 2015).

More recently, some studies have assessed job performance over longer periods of time. At a broad level, fluctuations in job engagement—which predicts job performance—have been conceptualized as being related to resources, such as social support and feedback, proactively influencing the job (termed "job crafting" in the literature), and daily recovery from stress (Bakker 2014). For example, a study following 288 participants over the course of three months found that job crafting at month two predicted job engagement and performance at month three, even after adjusting for month one measures (Tims et al. 2015). Another study measured job performance weekly for a month, finding that it was predicted by participants feeling "recovered" after the weekend (Binnewies et al. 2010). Weekend recovery was predicted by psychological detachment, relaxation, and mastery of experiences. The advantage of data that include longer timescales is that the resulting models are robust to within-person fluctuations in job performance, which could be due to changes in stress or outside life circumstances that happen at longer timescales (e.g., seasonal changes, having a child).

Notably, the literature cited previously primarily investigates self-reported measures as correlates of job performance. In contrast, studies of other, related constructs, such as cognitive ability, have used passive sensing devices to obtain

objectively measured correlates. For example, decline in performance by a group of 29 active-duty service members over 2 h of cognitive testing was significantly associated with electrodermal activity and two features extracted from recorded speech (Heaton et al. 2020). Additionally, heart rate variability features have also been used to develop personalized models able to predict declines in cognitive performance (Tsunoda et al. 2017).

In the work presented here, we combine the longer temporal scale of studies that measure job performance over the course of days, weeks, or months with more objective sensor data, a rarity in the literature. Specifically, using data from a long-term longitudinal study, we develop machine learning models that classify self-reported performance as a function of daily time series of heart rate and step counts taken from wearable sensors. We focus on a classification method that allows for the importance of multi-timescale features to emerge, which is distinct from much of the existing literature (e.g., Tims et al. 2015; Alessandri et al. 2018, but see Puig-Ribera et al. 2008), which tends to conceive of job performance as a function of single time-point measures (e.g., average response on a questionnaire), albeit taken repeatedly through time. Furthermore, much of the work on job performance focuses on identifying correlates through statistical models, which does not provide good information on potential out-of-sample prediction in the face of unseen data. We address this issue through *k*-fold cross validation. We argue that the approach taken here accounts for sources of variability in a way that previous work did not, and that this is key for making real-world decisions and building interventions that can prevent mid- and long-term performance decreases.

1.1 Why Heart Rate and Physical Activity?

Research has consistently found an interrelationship among biological, psychological, and social domains, suggesting that they can be conceived of as part of a single larger system (Ader and Cohen 1995). From a dynamic systems perspective, it may therefore be possible to detect aspects of the system based on the dynamics of a single or a small number of channels (Sugihara and May 1990; Chang et al. 2017).

Heart rate is a particularly promising candidate for detecting these complex interactions because a wealth of prior research has linked it to physical, social, and psychological outcomes. For example, patterns of heart rate variability have been linked to personality traits (Oveis et al. 2009; Kogan et al. 2014), social perceptions by others (Kogan et al. 2013), and stress responses (Berntson and Cacioppo 2004). Patterns of heart rate variability have also been linked to physical health in the form of predicting future myocardial infarction (Chattipakorn et al. 2007). Additionally,

analyses have found that changes in intraday heart rate and heart rate variability can be an indicator of specific periods of elevated stress, particularly when combined with information about physical movement (Verkuil et al. 2016).

This last point, that heart rate in combination with a measure of physical activity has particular predictive power, is potentially critical for understanding psychological outcomes. Patterns of heart rate assessed throughout the day are influenced by internal subjective states, such as stress and affect, but also heavily influenced by physical behavior, such as walking or typing (Vrijkotte et al. 2000; Rennie et al. 2003; Hjortskov et al. 2004; Thayer et al. 2012). Recent research has suggested that estimates of heart rate and heart rate variability that are statistically adjusted for levels of physical activity may give a more precise estimate of when an individual is experiencing a stressful episode (Verkuil et al. 2016; Brouwer et al. 2018; Brown et al. 2020). These methods attempt to disentangle metabolic and psychological influences on heart rate, creating an estimate of heart rate fluctuations that are due primarily to internal appraisals—such as experiencing a stimulus as threatening or stressful.

Research has long established that passive sensing of physiology and activity can provide insight into psychological states, but foundational work relied on smaller samples run in a laboratory setting. With the proliferation of inexpensive wearable technologies in recent years, it has become feasible to collect data at a larger scale. Steps and heart rate are among the most ubiquitously collected data streams in commercially available wearables, making them excellent candidates for the development of models that can be implemented widely and in combination. Thus, physical activity and heart rate meet both theoretical and practical considerations for use in predictive modeling efforts.

1.2 Secondary and Contextual Features

In addition to using heart rate and physical activity for prediction, this data set contained important contextual features likely to be related to workplace performance. The first is the day of the week. Work weeks have consistent cycles to them, and feeling recovered after a weekend versus fatigued at the end of a week has been related to changes in alertness and performance on information processing tasks (Wellens and Smith 2006).

Additionally, data was collected on social activity during the week. Participants had a mobile phone with the Electronically Activated Recorder (EAR) app installed (Mehl 2017). This app makes brief recordings of ambient sounds, which can then be coded to determine patterns of social activity during daily life. Social support has been identified as a predictor of job performance in previous research, and

EAR-coded speech has been related to general satisfaction with life (Bakker 2014; Milek et al. 2018).

In previous research, sound recordings were manually coded, but in this case, EAR files needed to be automatically processed to preserve anonymity. Additionally, participants were asked by their organization not to bring the device into certain meetings or areas of the office complex. Since this caused significant patterns of non-random missingness, the EAR data was not treated as a time series, but instead the average proportion of speech for the whole day was estimated from available data. This provides some context regarding whether an individual had a more social or more solitary day, which may help predict job performance. Day of the week and amount of socializing on a given day were used as additional predictors in modeling (as described later), to account for context.

1.3 Multi-Timescale Features and Time-Series Classification

Our goal in this work is to relate daily time series of passively sensed data from a wearable device to self-reported job performance. In the literature from psychology research, typical summary statistics of a given time series, such as the mean and standard deviation, are often correlated with an outcome of interest (e.g., Kuppens et al. 2010; Gruber et al. 2015; Sloan et al. 2017). In some less-common cases, nonlinear methods are used to generate features that can be correlated with these outcomes (e.g., Danvers et al. 2020, Likens et al., 2018). The major advantage of these approaches is interpretability: If the summary statistics or nonlinear features are well understood and have theoretical meaning, such methods can assist in further theory development.

The major disadvantage of the approach typical of psychology is a lack of flexibility in representation on the part of the model, and therefore, poorer out-of-sample performance. There are infinite ways of quantitatively characterizing a given time series, and it may be that there is an important feature of the time series that relates to the outcome of interest that is nonlinear, happening at multiple timescales, and far outside of the set of typical features (e.g., means and standard deviations) considered. One could expect this in the case of job performance because people vary both day to day and throughout the day in their level of engagement and productivity at work based on the interplay of many factors, both externally determined (e.g., deadlines, events) and internal (e.g., chronic stress, sleep disturbances).

In the machine learning literature, our goal of relating sensed data to job performance would be defined as an example of so-called time-series classification—that is, finding a function that matches one or more time series to a

categorical outcome, also called a class or label. The function is typically allowed to be arbitrarily complex and is learned by optimizing for classification performance over interpretability. In this work, we use Random Convolutional Kernel Transform (ROCKET), an algorithm that generates a very large number of multi-timescale features from input time series and then matches them to labels (Dempster et al. 2020).

2. Methods, Assumptions, and Procedures

2.1 Participants

Participants were individuals who participated in a 60-day longitudinal study at a Silicon Valley office. Data from 212 participants who had completed at least one outcome measure were considered. On average, 26 days per person were considered for the study.

2.2 Procedures

Participants were instructed to wear sensors while at work during the 60-day period when they were enrolled in the study. Enrollment in the study was on a rolling basis, so participants did not all participate during the same set of days. No specific instructions or behavioral manipulations were provided to participants, so that data represented a naturalistic, ecologically valid sampling of their daily life in the office context.

2.3 Measures

Heart Rate. Participants wore a Fitbit Charge 4 device throughout the day. The Fitbit device captures heart rate through photoplethysmography (PPG) and outputs heart rate data through internal processing algorithms. Raw PPG data was not stored, so the estimated heart rate provided by the Fitbit processing algorithm was used. Some participants wore their Fitbit devices for more than their period at work, but only data from 7 AM to 4 PM (approximating work hours for the full day until the questionnaire was sent out) was used. Mean and standard deviation of heart rate across 5-min samples were used as input time series to ROCKET.

Physical Activity. Step counts recorded by the Fitbit Charge 4 device were taken as an index of physical activity. As with heart rate, data from 7 AM to 4 PM was used and both mean and standard deviation across 5-min segments were included as features.

Social Activity. Participants were given Android smartphones with the EAR app downloaded and programmed. This app recorded 10-s segments of ambient audio every 5 min. The raw audio was processed using the Google Audioset algorithm, a machine learning algorithm that was trained to identify specific sound categories from 8 million YouTube audio clips. The sound category "speech" from each processed audio file was saved and used as an indicator of social activity.

Day of the Week. The day of the week on which measures were taken was saved as a contextual feature for use in the model.

Job Performance. Participants responded to a survey sent to them every third day at 4 PM. The survey asked them to reflect on their job performance "today" and indicate their agreement with a series of statements about job performance. Based on initial observation, we found that responses to many items had very low variability. These low variability items were excluded.

2.3.1 Subjective Job Performance

Two scales were administered that assessed a participant's subjective job performance. The daily versions of these scales were adapted from well-validated industrial/organizational psychology instruments.

In-Role Behavior (IRB). Items were adapted from the IRB used in Williams and Anderson (1991). Participants responded on a seven-point Likert scale, from "strongly disagree" to "strongly agree". There were seven items included in the full scale, but for three items, the most socially desirable response was given over 60% of the time. The low variability in responses to these items led us to remove them from the scale. The items removed were "Performed tasks that are expected of you", "neglected aspects of the job you are obligated to perform", and "failed to perform essential duties". The last two items are reverse coded when included in the scale. The following items were retained:

1) Adequately completed your assigned duties
2) Fulfilled responsibilities specified in your job description
3) Met formal performance requirements of your job
4) Engaged in activities that will directly affect your performance evaluation

Individual Task Performance (ITP). Items were adapted from Griffin et al. (2007). Participants responded on a five-point Likert scale, from "strongly disagree" to "strongly agree" to a series of items indicating ITP. The following items were included:

1) Carried out the core parts of your job well
2) Completed your core tasks well using the standard procedures
3) Ensured your tasks were completed properly

To check that items from these scales did capture distinct constructs, parallel analysis was used to determine the number of factors that should be extracted in an exploratory factor analysis. Parallel analysis uses resampled data to empirically determine expected null eigenvalues for each number of factors in an exploratory factor analysis and suggests that any factors with eigenvalues above this level be retained. Parallel analysis suggested retaining two factors, and exploratory factor analysis using two factors found that the items from the two scales loaded highly on separate factors with minimal cross-loadings. The two factors were moderately correlated ($r = 0.67$). Therefore, separate IRB and ITP scores were estimated from this exploratory factor analysis.

Histograms of IRB and ITP factor scores were plotted. Based on a visual inspection of the histograms, we identified natural cut points. Outcomes were therefore split into categories at each of these cut points, which created low, medium, and high groupings for the IRB and ITP scales. These cut points are plotted in Fig. 1.

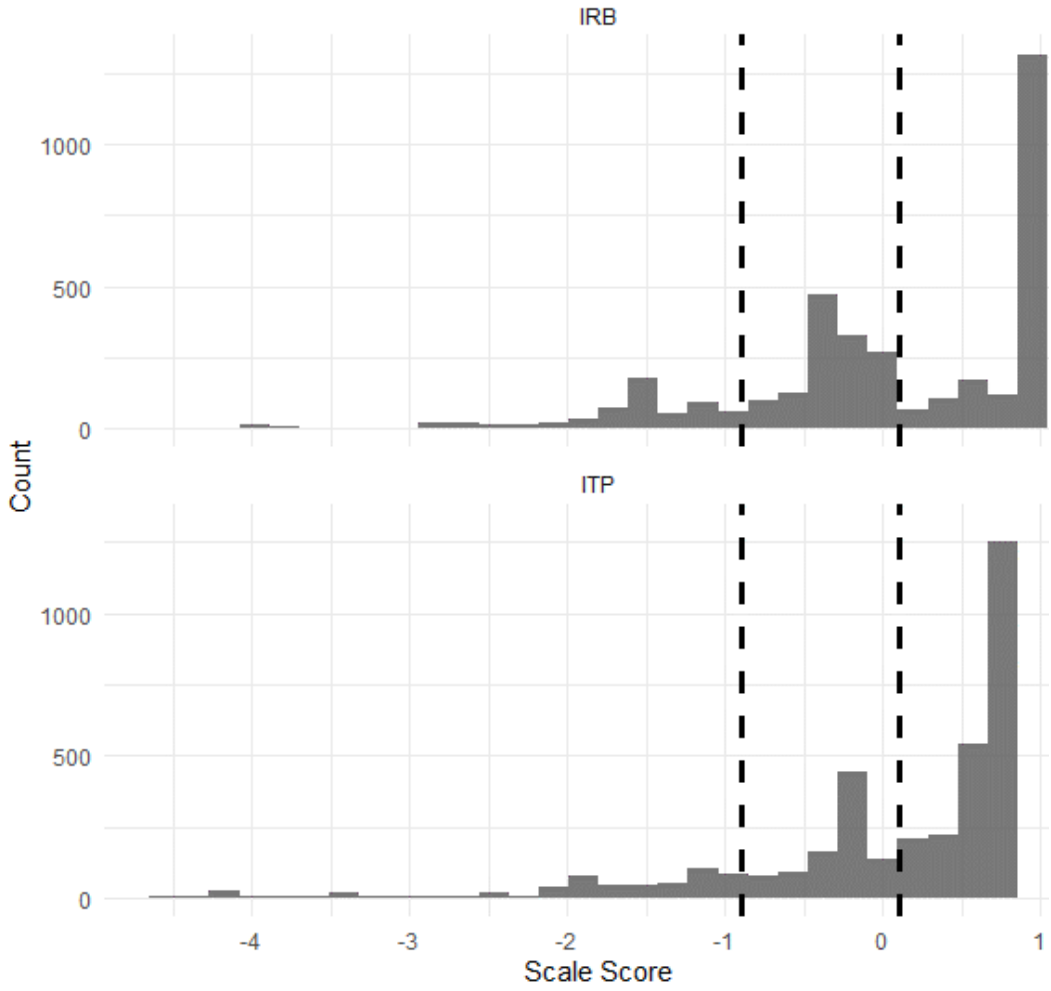

Fig. 1 Histograms of subjective job performance scores with cut points

2.3.2 Objective Job Performance

Participants were also asked about performing specific behaviors on the day of assessment, providing a more-objective assessment of job performance.

Specific Behaviors. Participants indicated whether they did or did not perform specific job-related behaviors. These behaviors were drawn from scales for Organizational Citizenship Behavior (OCB; Fox et al. 2012) and Counterproductive Workplace Behavior (CWB; Bennett and Robinson 2000). In almost all cases, the socially desirable response was given over 80% of the time, suggesting that they did not track regular changes in work performance well. However, one OCB and one CWB item had a variability lower than 66% (e.g., the desirable response was chosen less than 66% of the time). These two items were retained for modeling. They are the following:

- Volunteered to do something that was not required. (OCB; 59% yes)

- Spent time on tasks unrelated to work. (CWB; 61% no)

The full list of items is provided in the Appendix.

2.4 Analysis

Our entire data set comprised 3693 input–output pairs, where a given input was represented as four time series (mean and a standard deviation of 5-min bins for heart rate and activity) of 108 samples. Outputs took two forms: two classes for the objective job outcomes "volunteered to do something not required" (yes/no) and "did unrelated work" (yes/no), and three classes (low/medium/high) for IRB and ITP. Note the same inputs were used for each different output in separate models.

Initially, 20% of the data set (pseudo-randomly selected such that the cases were balanced on the variable "volunteered") was held out as a test set. Balancing was done on only one variable so that the test data would be the same when training and evaluating all models. When training models on the remaining 80% of the data, five-fold cross-validation was used for grid-search-based hyperparameter tuning, where folds were made such that classes were balanced on instances of each value of the associated outcome. This balancing was done separately for each outcome.

As mentioned, our primary modeling approach uses ROCKET for times-series classification. ROCKET includes two steps. First, a very large number of random kernels are generated and applied to the input time series, thereby dramatically expanding the feature space. A given kernel is defined by its length, weight and bias values, padding, and dilation. This approach is similar to what is typical of feature extraction via convolutional layers in convolutional neural networks (CNNs); however, in CNNs, kernels are learned from the data as opposed to randomly generated. In the second step, the expanded feature space is fed to a classification algorithm (e.g., ridge regression in the original conceptualization). Critically, these two steps have been shown to result in out-of-sample classification performance at or above other state-of-the-art methods but with significantly lower computational costs (Ruiz et al. 2021). This method is particularly well suited to our data, given the relatively few instances (at least compared to data in deep learning contexts) and the hypothesized importance of multi-timescale features. Specifically, kernel dilation is a kind of "stretching" operation that, along with changing kernel size, can capture information from different timescales. In development, this multiscale information was found to be important to ROCKET's predictive accuracy.

ROCKET includes several hyperparameters. First is the number of kernels, where more kernels yield linearly increasing accuracy, but increasing slow-down in

training. The default, based on 85 time series tested by Dempster et al. (2020), is 10,000; increases beyond this point did not significantly improve accuracy. In the analysis presented here, only sets of 10,000 kernels were considered. The second hyperparameter is the set of sizes possible for random kernels. Two groupings of kernel sizes were considered. In the small group, kernels could be of size 3, 7, 15, or 32 points. In the large group, kernels could be of size 3, 9, 27, or 79 points. The longer kernel sizes were explored due to the longer length of the time series used here (i.e., 108) relative to many of the examples used to validate the ROCKET algorithm.

We assessed two classification algorithms for the second step of ROCKET. The first, ridge regression, which was used in the original version of ROCKET, adapts a penalized regression model for binary or multi-class classification. The ridge classifier also has a tunable hyperparameter in its penalty parameter. A series of 20 evenly spaced penalty values ranging from 0.5 to 3 was used. The second, random forests, is an ensemble method that combines decision trees. For random forests, the number of trees and the number of features to consider at each decision point are hyperparameters and were tuned using a grid search. Values for the number of trees considered were 500, 1000, and 2000. Values for the number of features to consider were 6, 36, and 60.

All ROCKET models were compared to baseline models where ridge regression and random forest models were fed the average of the input time-series features: heart rate, heart rate variability (average of the standard deviation series), physical activity, variability in physical activity (average of the standard deviation series), and average amount of speech during the day. Hyperparameters for these baseline models were tuned similarly to their analogs in our implementation of ROCKET. Note, however, that the number of features to consider for random forest baseline models could not include 36 and 60, since there were not sufficient features for these options.

The performance of all models was assessed using the Matthew's Correlation Coefficient (MCC; Gorodkin 2004) for K classes:

$$MCC = \frac{c \times s - \sum_k^K p_k \times t_k}{\sqrt{(s^2 - \sum_k^K p_k^2) \times (s^2 - \sum_k^K t_k^2)}}. \quad (1)$$

where c is the number of instances correctly predicted, s is the total number of instances in the data set, t_k is the number of times class k occurred, and p_k is the number of times class k was predicted. Perfect prediction is represented by MCC = 1, whereas average random guessing is represented by MCC = 0. The

minimum value of MCC for two classes is –1 and minimum value for multi-class MCC is somewhere between 0 and –1 and depends on the data in question.

3. Results and Discussion

Demographics. Summaries of demographic information for participants, designed by the study group to reduce identifiability, is presented in Table 1.

Table 1 Summary of participant demographics

Demographics	Response options/bins	N	%
Gender identity	Female	84	40%
	Male	128	60%
Age	18-34	101	48%
	35-44	40	19%
	45+	71	33%
Education Level	Some HS – Some college	15	7%
	College degree – Some graduate	130	61%
	Master's degree – doctoral degree	67	32%
Supervise others	Yes	53	25%
	No	159	75%
Time with employer	< 10 years	109	51%
	10+ years	103	49%

Classification performance for all models is presented in Table 2.

Table 2 Classification performance on the test data for all models

Outcome	Classifier	Baseline: MCC	ROCKET small: MCC (%Δ)	ROCKET large: MCC (%Δ)
IRB	Ridge	0.045	0.057 (27%)	**0.062 (38%)**
IRB	RF	0.066	0.077 (17%)	**0.11 (67%)**
ITP	Ridge	0.06	0.065 (8%)	**0.068 (13%)**
ITP	RF	0.049	**0.099 (102%)**	0.064 (31%)
Volunteer	Ridge	0.096	**0.142 (48%)**	0.135 (41%)
Volunteer	RF	-0.013	0.01 (177%)	**0.031 (339%)**
Unrelated	Ridge	0.028	**0.093 (232%)**	0.087 (211%)
Unrelated	RF	0.042	**0.061 (45%)**	0.048 (14%)

Note: Volunteer = volunteering to do additional work; Unrelated = doing work unrelated to one's job at work; Ridge = ridge regression classifier; RF = random forests classifier; ROCKET small = small kernel sizes (3, 7, 15, 32); ROCKET large = large kernel sizes (3, 9, 27, 79); %**Δ** = percentage change of ROCKET model performance from baseline performance. When MCC was negative, the proportion change value was divided by the absolute value of the baseline MCC.

Results from the baseline models indicate that the mean values of inputs across a day, combined with the day of the week, can provide better than random performance for all outcomes. When the ROCKET algorithm was used, predictions increased, and in some cases substantially, for all outcomes. The best-performing baseline model was random forest applied to IRB at MCC = 0.066, whereas the

best performance for a ROCKET model for IRB was MCC = 0.110, an increase of 67%. This same pattern held across all outcomes with improvements of 65%, 48%, and 121% for ITP, volunteer, and unrelated, respectfully.

For all variables except IRB, the ridge regression version of ROCKET was better than the random forest version. This likely reflects the dimensionality of the data. Ridge regression automatically accounts for all the features—meaning the maximum and number of positive predictive values for each kernel, plus baseline features—in its additive, weighted model. However, random forest models select a random group of predictors at each step, potentially missing some that might be useful. More comprehensively sampling the space of predictors through random forests would require growing a number of decision trees comparable to the number of kernels, which is not feasible computationally (and removes the low computational cost that is one of the primary benefits of ROCKET). Instead, random forest baseline models tended to do better than the ridge regression baseline models, likely because of the smaller set of features. This is expected, as this smaller set of features can be more fully explored in all its combinations.

Although using ROCKET substantially improved classification performance, overall performance remains modest. The MCC values for the final models can be compared to Pearson's correlation coefficient, as MCC is a statistic trying to capture the same information from a contingency table. From that perspective, all of the final MCC values are comparable to what is considered a small effect in psychology (MCC = 0.10, analog to $r = 0.10$) (Funder and Ozer 2019). Small effects can have substantial importance in theory development and are common in behavioral science. However, they leave substantial room for improvement, potentially through the use of a broader set of predictors.

4. Conclusions

The ROCKET algorithm improved classification performance of all job performance metrics by a substantial amount—ranging from 48% to 121%. The substantial improvements are in line with previous research applying the ROCKET algorithm to a variety of time-series classification tasks (Dempster et al. 2020). The problem addressed here is one that has long been difficult for psychologists: predicting subjective self-report data from objective measures. This machine learning approach has made significant progress on increasing predictive accuracy, but not yet solved this problem.

Volunteering to do something that was not required on a given day was the most accurately predicted outcome. Volunteering reflects capacity (e.g., having the time and energy to do something additional) as well as motivation (e.g., having the desire

to help colleagues and the organization). Changes in heart rate can reflect stress and engagement, which can influence capacity and motivation. High levels of stress reduces capacity to take on more work, while high levels of task engagement may reflect motivation to complete workplace projects. This specific behavior appears to be most closely connected with the information picked up by the sensors used in this study.

IRB was the next-most-accurate model. This variable reflects a self-reported assessment that the individual was performing their job duties adequately. Accounting for changes in activity and heart rate may be able to provide more insight into this behavior due to the way these reflect workplace engagement. Low levels of physical activity may reflect time spent in "heads-down work", while high levels of heart rate variability may reflect focused engagement with a task.

Overall, the most accurate models for all outcomes were comparable. Beyond volunteering behavior, which was slightly higher (MCC = 0.142), accuracy for all other predictors was around MCC = 0.10 (for IRB, MCC = 0.110; for ITP, MCC = 0.099; for unrelated work, MCC = 0.093).

Classification performance, as assessed by MCC, was still modest overall. This is in line with findings from previous large-scale projects attempting to predict daily outcomes from this data (Mattingly et al. 2019). Translating low-level features, like heart rate and physical activity, into indicators of a higher level construct, like workplace performance, remains challenging.

5. Future Directions and Limitations

The data collected here, while representing a large and relatively diverse sample as compared to typical behavioral science studies, only samples from a single office with employees at a large company doing white collar work. It is likely that patterns of heart rate related to job performance would be significantly different if employees doing different types of work, such as manual labor, customer service, or deployment in the field, were studied. Researchers collecting data in the current work environment would also do well to sample from employees working from home, as this is an increasingly common work context. Additionally, patterns of prediction may differ across cultural or ethnic groups, potentially based on learned interpretations of physiological signals or underlying levels of life stress.

6. References

Ader R, Cohen N. Psychoneuroimmunology: interactions between the nervous system and the immune system. The Lancet. 1995;345(8942):99–103.

Alessandri G, Consiglio C, Luthans F, Borgogni L. Testing a dynamic model of the impact of psychological capital on work engagement and job performance. Career Development International; 2018.

Bakker AB. Daily fluctuations in work engagement. European Psychologist. 2014.

Bennett RJ, Robinson SL. Development of a measure of workplace deviance. J Appl Psychol. 2000;85(3):349.

Berntson GG, Cacioppo JT.. Heart rate variability: stress and psychiatric conditions. Dynamic Electrocardiog. 2004;41(2):57–64.

Binnewies C, Sonnentag S, Mojza EJ. Recovery during the weekend and fluctuations in weekly job performance: a week-level study examining intra-individual relationships. J Occupat Org Psychol. 2010;83(2):419–441.

Brouwer AM, van Dam E, Van Erp JB, Spangler DP, Brooks JR. Improving real-life estimates of emotion based on heart rate: a perspective on taking metabolic heart rate into account. Front Hum Neurosciv. 2018;12(284).

Brown SBRE, Brosschot JF, Versluis A, Thayer JF, Verkuil B. Assessing new methods to optimally detect episodes of non-metabolic heart rate variability reduction as an indicator of psychological stress in everyday life: a thorough evaluation of six methods. Front Neurosci. 2020 Oct 22;14:564123. doi: 10.3389/fnins.2020.564123.

Chang CW, Ushio M, Hsieh CH. Empirical dynamic modeling for beginners. Ecol Res. 2017;32(6):785–796.

Chattipakorn N, Incharoen T, Kanlop N, Chattipakorn S. Heart rate variability in myocardial infarction and heart failure. Int J Cardiol. 2007;120(3):289–296.

Danvers AF, Sbarra DA, Mehl MR. Understanding personality through patterns of daily socializing: applying recurrence quantification analysis to naturalistically observed intensive longitudinal social interaction data. European Journal of Personality. 2020;34(5):777–793.

Dempster A, Petitjean F, Webb GI. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. Data Mining Knowl Disc. 2020;34(5):1454–1495.

Fox S, Spector PE, Goh A, Bruursema K, Kessler SR. The deviant citizen: measuring potential positive relations between counterproductive work behaviour and organizational citizenship behaviour. J Occupat Org Psychol. 2012;85(1):199–220.

Funder DC, Ozer DJ. Evaluating effect size in psychological research: sense and nonsense. Adv Meth Pract Psychol Sci. 2019;2(2):156–168.

Gorodkin J. Comparing two K-category assignments by a K-category correlation coefficient. Comput Biol Chem. 2004;28(5–6):367–374.

Griffin MA, Neal A, Parker SK. A new model of work role performance: positive behavior in uncertain and interdependent contexts. Acad Manage J. 2007;50(2):327–347.

Gruber J, Mennin DS, Fields A, Purcell A, Murray G. Heart rate variability as a potential indicator of positive valence system disturbance: a proof of concept investigation. Int J Psychophys. 2015;98(2):240–248.

Heaton KJ, Williamson JR, Lammert AC, Finkelstein KR, Haven CC, Sturim D, ... Quatieri TF. Predicting changes in performance due to cognitive fatigue: a multimodal approach based on speech motor coordination and electrodermal activity. Clinical Neuropsychol. 2020;34(6):1190–1214.

Hjortskov N, Rissén D, Blangsted AK, Fallentin N, Lundberg U, Søgaard K. The effect of mental stress on heart rate variability and blood pressure during computer work. Euro J Appl Phys. 2004;92(1):84–89.

Hurtz GM, Donovan JJ. Personality and job performance: the Big Five revisited. J Appl Psychol. 2000;85(6):869.

Joseph DL, Jin J, Newman DA, O'Boyle EH. Why does self-reported emotional intelligence predict job performance? A meta-analytic investigation of mixed EI. J Appl Psychol. 2015;100(2):298.

Kogan A, Gruber J, Shallcross AJ, Ford BQ, Mauss IB. Too much of a good thing? cardiac vagal tone's nonlinear relationship with well-being. Emotion. 2013;13(4):599.

Kogan A, Oveis C, Carr EW, Gruber J, Mauss IB, Shallcross A, Impett EA, van der Lowe I, Hui B, Cheng C, Keltner D. Vagal activity is quadratically related to prosocial traits, prosocial emotions, and observer perceptions of prosociality. J Personality Soc Psychol. 2014;107(6):1051.

Kuppens P, Allen NB, Sheeber LB. Emotional inertia and psychological maladjustment. Psychol Sci. 2010;21(7):984–991.

Likens AD, McCarthy KS, Allen LK, McNamara DS. Recurrence Quantification Analysis as a method for studying text comprehension dynamics. Proceedings of the 8th International Conference on Learning Analytics and Knowledge; 2018. p. 111–120.

Mattingly SM, Gregg JM, Audia P, Bayraktaroglu AE, Campbell AT, Chawla NV, ..., Striegel A. The Tesserae project: large-scale, longitudinal, in situ, multimodal sensing of information workers. Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems; 2019, May. p. 1–8.

McDaniel MA, Morgeson FP, Finnegan EB, Campion MA, Braverman EP. Use of situational judgment tests to predict job performance: a clarification of the literature. J Appl Psychol. 2001;86(4):730.

Mehl MR. The Electronically Activated Recorder (EAR) a method for the naturalistic observation of daily social behavior. Current Directions Psychol Sci. 2017;26(2):184–190.

Milek, A, Butler EA, Tackman AM, Kaplan DM, Raison CL, Sbarra DA, Vazire S, Mehl MR. "Eavesdropping on happiness" revisited: a pooled, multisample replication of the association between life satisfaction and observed daily conversation quantity and quality. Psychological Science. 2018;29(9):1451–1462.

Oveis C, Cohen AB, Gruber J, Shiota MN, Haidt J, Keltner D. Resting respiratory sinus arrhythmia is associated with tonic positive emotionality. Emotion. 2009 Apr;9(2):265-270. doi: 10.1037/a0015383.

Puig-Ribera A, McKenna J, Gilson N, Brown WJ. Change in work day step counts, wellbeing and job performance in Catalan university employees: a randomised controlled trial. Promotion Education. 2008;15(4):11–16.

Rennie KL, Hemingway H, Kumari M, Brunner E, Malik M, Marmot M. Effects of moderate and vigorous physical activity on heart rate variability in a British study of civil servants. American Journal of Epidemiology. 2003 Jul 15;158(2):135–143, https://doi.org/10.1093/aje/kwg120.

Ruiz AP, Flynn M, Large J, Middlehurst M, Bagnall A. The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mining Knowl Discov. 2021;35(2):401–449.

Salgado JF. Predicting job performance using FFM and non-FFM personality measures. J Occupat Org Psychol. 2003;76(3):323–346.

Sloan RP, Schwarz E, McKinley PS, Weinstein M, Love G, Ryff C, Mroczek D, Choo T-H, Lee S, Seeman T. Vagally-mediated heart rate variability and indices of well-being: Results of a nationally representative study. Health Psychol. 2017;36(1):73.

Sugihara G, May RM. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. Nature. 1990;344(6268):734–741.

Thayer JF, Åhs F, Fredrikson M, Sollers JJ 3rd, Wager TD. A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. Neurosci Biobehav Rev. 2012;36(2):747–756.

Tims M, Bakker AB, Derks D. Job crafting and job performance: a longitudinal study. Euro J Work Org Psychol. 2015;24(6):914–928.

Tracey JB, Sturman MC, Tews MJ. Ability versus personality: factors that predict employee job performance. Cornell Hotel Restaurant Admin Quarterly. 2007;48(3):313–322.

Tsunoda K, Chiba A, Yoshida K, Watanabe T, Mizuno O. Predicting changes in cognitive performance using heart rate variability. IEICE Trans Info Syst. 2017;100(10):2411–2419.

Verkuil B, Brosschot JF, Tollenaar MS, Lane RD, Thayer JF. Prolonged non-metabolic heart rate variability reduction as a physiological marker of psychological stress in daily life. Annals Behav Med. 2016;50(5):704–714.

Vrijkotte TG, Van Doornen LJ, De Geus EJ. Effects of work stress on ambulatory blood pressure, heart rate, and heart rate variability. Hypertension. 2000;35(4):880–886.

Wellens BT, Smith AP. Combined workplace stressors and their relationship with mood, physiology, and performance. Work Stress. 2006;20(3):245–258.

Williams LJ, Anderson SE. Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. J Manage. 1991;17(3):601–617.

Appendix. All Job Performance Questions

Original IRB Items

The following questions concern your perceptions about your job performance *today*.

Please indicate your level of agreement with whether you...

1) Adequately completed your assigned duties
2) Fulfilled responsibilities specified in your job description
3) Performed tasks that are expected of you
4) Met formal performance requirements of your job
5) Engaged in activities that will directly affect your performance evaluation
6) Neglected aspects of the job you are obligated to perform [R]
7) Failed to perform essential duties [R]

Response scale ranges from 1 (Strongly disagree) to 7 (Strongly agree).

[R] = Reverse scored.

Original ITP Items

This inventory pertains to behaviors you perform on your job. Please indicate how often you carried out these three behaviors *today*.

1) Carried out the core parts of your job well
2) Completed your core tasks well using the standard procedures
3) Ensured your tasks were completed properly

Response scale: 1 (Very little); 2 (Somewhat); 3 (Moderately); 4 (Considerably); 5 (A great deal).

Original OCB and CWB Items

Today, I…

1) Went out of my way to be a good employee. (OCB)
2) Was respectful of other people's needs. (OCB)
3) Displayed loyalty to my organization. (OCB)
4) Praised or encouraged someone. (OCB)
5) Volunteered to do something that was not required. (OCB)

6) Showed genuine concern for others. (OCB)

7) Tried to uphold the values of my organization. (OCB)

8) Tried to be considerate to others. (OCB)

9) Spent time on tasks unrelated to work. (CWB)

10) Gossiped about people at my organization. (CWB)

11) Did not work to the best of my ability. (CWB)

12) Said or did something that was unpleasant. (CWB)

13) Did not fully comply with a supervisor's instructions. (CWB)

14) Behaved in an unfriendly manner. (CWB)

15) Spoke poorly about my organization to others. (CWB)

16) Talked badly about people behind their backs. (CWB)

(OCB) = OCB items; (CWB) = CWB items.

Response scale is Yes/No.

List of Symbols, Abbreviations, and Acronyms

ARL	Army Research Laboratory
CNN	convolutional neural networks
CWB	Counterproductive Workplace Behavior
DEVCOM	US Army Combat Capabilities Development Command
EAR	Electronically Activated Recorder
IRB	in role behavior
ITP	individual task proficiency
MCC	Matthew's Correlation Coefficient
OCB	Organizational Citizenship Behavior
PPG	photoplethysmography
RF	random forests
ROCKET	Random Convolutional Kernel Transform

1 (PDF) DEFENSE TECHNICAL INFORMATION CTR DTIC OCA

1 (PDF) DEVCOM ARL FCDD RLD DCI TECH LIB

1 (PDF) DEVCOM ARL FCDD RLH B T DAVIS BLDG 5400 RM C242 REDSTONE ARSENAL AL 35898-7290

1 (PDF) DEVCOM ARL FCDD HSI J THOMAS 6662 GUNNER CIRCLE ABERDEEN PROVING GROUND MD 21005-5201

1 (PDF) USN ONR ONR CODE 341 J TANGNEY 875 N RANDOLPH STREET BLDG 87 ARLINGTON VA 22203-1986

1 (PDF) USA NSRDEC RDNS D D TAMILIO 10 GENERAL GREENE AVE NATICK MA 01760-2642

1 (PDF) OSD OUSD ATL HPT&B B PETRO 4800 MARK CENTER DRIVE SUITE 17E08 ALEXANDRIA VA 22350

ABERDEEN PROVING GROUND

14 (PDF) DEVCOM ARL FCDD RLH J LANE Y CHEN P FRANASZCZUK A MARATHE K MCDOWELL K OIE FCDD RLH F J GASTON (A) FCDD RLH FA A DECOSTANZA A DANVERS E CARTER FCDD RLH FB D BOOTHE (A) FCDD RLH FC K COX (A) FCDD RLH FD A FOOTS (A) FCDD RLH FE D HEADLEY