**Abstract**

**Data Collection and Preprocessing**

This study introduces a data oriented approach to investigate the research question described earlier.

The first step, is data collection. There are predominantly two different types of data that are collected in this study. First is, sensor data pertaining to time stamped movement, distance, back angle, and rate of change of acceleration. Second is, self-reported exertion data based on Borg rating [1]. The sensor data is gathered through sensors mounted on the subjects. Data is captured through API and stored in excel files for pre-processing. These self-reported Borg rates are collected by interrupting the subjects every fifteen min. Data for five different variables is recorded every 2 minutes over the task duration in a time series format. Total task time is 167 minutes with 15 subjects employed for this study. A large dataset for each subject is collected in this experiment and every two minutes a data point is created. The following table demonstrates interpretation of the Borg’s values. The larger values of Borg are associated to higher fatigue levels:

|  |  |  |  |
| --- | --- | --- | --- |
| RPE | Level of fatigue | Level of exertion | verbal anchor |
| 6 | 1- Low | no exertion | I am not tired (similar to resting) |
| 7 |  |
| 7.5 | extremely light |
| 8 |  | I am not tired (similar to walking) |
| 9 | very light |
| 10 |  |
| 11 | light | I feel fine to continue |
| 12 | 2- Medium |  |
| 13 | somewhat light |  |
| 14 |  | I am getting tired, but I can continue. |
| 15 | 3- High | hard (heavy) |
| 16 |  |
| 17 | 4- Very High | very hard | I am very tired, I have to push myself to continue. |
| 18 |  |
| 19 | extremely hard | This is one of the hardest things I have done. |
| 20 | maximal exertion |

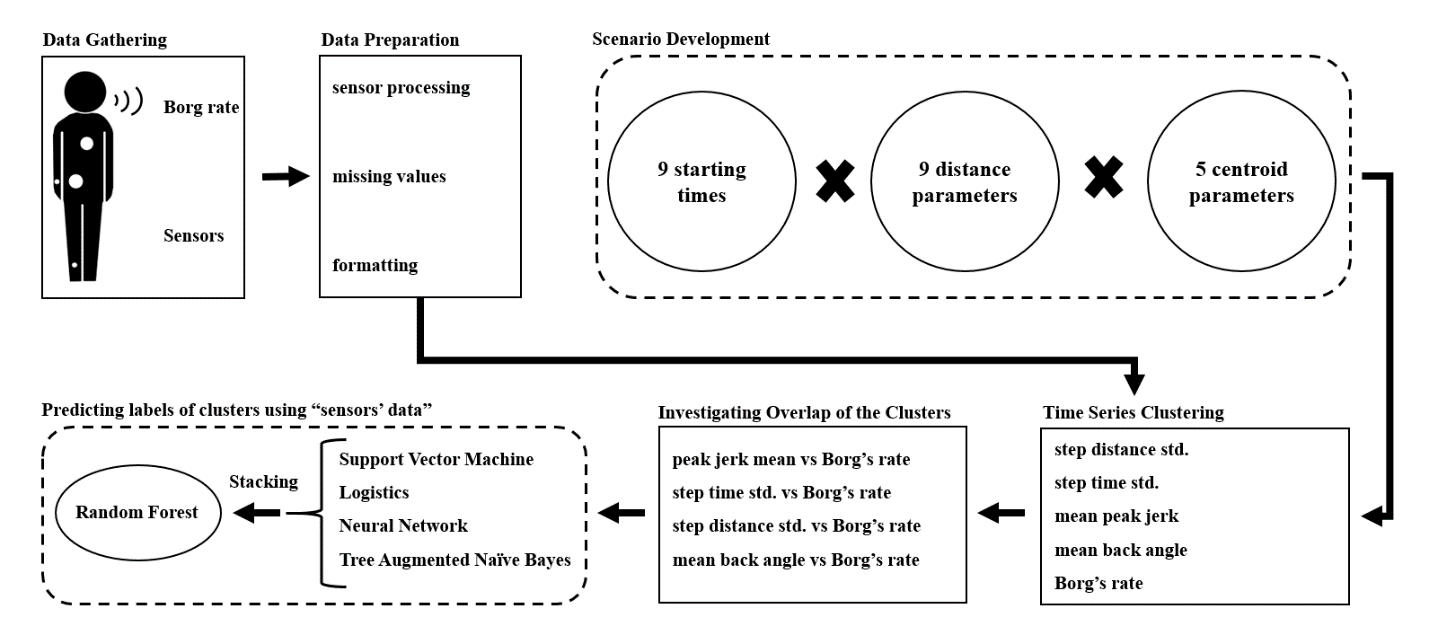
Figure : Borg's values interpretation (reference needed?!!! I put from Dr. Cavuto email)

Data collected is then preprocessed for missing values. No imputation is performed on the raw data and the longest time series that has the complete data for a particular subject is considered for further investigation. Borg’s value represents the subjects’ report about their fatigue’s level. The four candidate variables (objective measures) of: “standard deviation of time”, “standard deviation of the length”, “Average of Peaks in Jerk Values”, and “Average of the Back Angle” are investigated to compare their similarity with the Borg’s value in terms of categorizing subjects. Performance of 15 subjects is recorded over all the 5 variables. The Borg’s value is recorded by asking from the subjects and recording their idea about their own fatigue. However, other variables are recorded through the sensors mounted on the subjects. The subjects are categorized to two clusters by employing partitional time series clustering [2, 3] over the time series. Then similarity of the produced clusters is compared.

After removing the missing time points, a time series of 50 data points is considered for analysis. After data preparation and cleaning process, similarity of sensor data and Borg ratings from self-reports is compared for performance assessment. The time series of sensors’ yields and the “reported Borg ratings” are utilized to categorize the subjects into two clusters.

Partitional time series algorithm is utilized for performing time series clustering. We then utilized a heuristic approach that uses nine distance parameters and five centroid parameters to identify the best three models that have the highest accordance with the “reported Borg rating”. Finally, we study the predictability of clusters developed by sensors data in the top five scenarios of each starting time cases. The top five labels are ensemble by majority voting approach. Then resulted labels are predicted.

The following figure summarized analysis methodology adopted in this study:



**Model Development**

A major goal of this study is developing an objective measure for understanding perceived level of fatigue using IoT data. Therefore, it enables investigation of fatigue without interfering with routine tasks or interrupting the workers. Use of IoT devices for capturing activity data provides an objective measure for assessing injury risk or fatigue-levels of an employee. This could eventually lead to better assigning for job routines based on the individual performance. We assess the alignment of these objective measures that are derived from IoT data with an employee’s subjective report about fatigue. This could lead to eventual replacement of direct inquiry from an employee while the task is still underway.

***Partitional time series clustering***: In this study, we use Partitional time series clustering for categorizing subjects into two groups based on their perceived fatigue and other candidate variables collected from IoT devices over the experiment duration. This approach to clustering is a type of optimization problem that maximizes inter-cluster distance while minimizes the intra-cluster distance. However, this method is a time extensive heuristic approach that enumerates all possibilities [2]. Therefore, searching a smaller portion of data through iterative greedy descent approach is adopted instead. In partitional time series clustering, k time series are randomly selected and considered as k centroids. Distance of other time series to the centroids are calculated and then assigned to the nearest time series. In the next step, a prototyping function [3] updates the centroids of each clusters. Centroids and distances are then updated through an iterative procedure until a certain number of iterations are reached or no change in the subjects’ labels is observed. This clustering algorithm is stochastic in nature and, therefore, a different random start might lead to a different local optima [3]. Therefore, considering different random seeds and selecting the best results is a common practice [3]. We adopted this approach while analyzing the data.

***Distance calculation method***: This study investigates nine different distance measure to calculate dissimilarity between the two time-series. The following equations demonstrate general distance calculation [3]:

(1)

where: x , y are time series

(1)

These measures are: dynamic time wrapping (dtw), dtw with (dtw2), fast version of dtw (), dtwwith proxy and Lemire's [4] lower bound () , dtw with Keogh's [5] lower bound (lbk), dtw with Lemire's [4] lower bound (lbi), shape-based distance (sbd), global alignment kernels (gak), and soft\_dtw (sdtw). dtw tries to calculate optimum wrapping distance between the candidate time series. This dynamic programming algorithm is developed to overcome some of constrains inherent in the Euclidian distance based computations [3, 6, 7]. However, such approaches are computationally extensive and prone to implementation bias [3]. This algorithm is executed in two steps. In the first step, a local cost matrix (LCM) is derived with n (data points of time series 1) x m (data points of time series 2). Then dtw searches a shortest path that minimizes alignment between elements of the time series. Distance parameter of dtw2, dtw\_basic, and dtw\_lb are special formations of the dtw. In dtw2, the norm parameter of the second step has the power of 2. dtw\_basic is the faster version of dtw algorithm but is mostly similar in functionality to dtw. Lower bounds (LBs) are developed to guarantee a distance between the time series that is less than or equal to the dtw distance [3]. Two of the proposed LBs are lbk [5] and lbi [4]. dtw\_lb leverages lbi [4] over the nearest points for improving distance calculations in dtw. The shape based distance (sbd) was developed as a subset of k-shape clustering [8], which is a faster algorithm compared to dtw. Sbd computes the cross-correlation with coefficient normalization (NCC) between the two time-series by convolving the two series [3]. The global alignment kernel (gak) computes the exponentiated soft-minimum as the cost of all possible alignments [3]. So the resulted similarities quantities is calculated in a more coherent approach [3]. Soft dtw (sdtw) regularizes the dtw algorithm through soothing [3, 9]. This algorithm provides gradient function (for centroid calculations) in a more efficient way [3]. This study utilizes all the measure to identify the best performing measures.

***Centroid calculation method***: In this study, we use 5 different approaches ("mean", "median", "shape", "DTW barycenter averaging", and "Partition around medoids") for centroid calculations. Mean and median are Euclidian distance parameters. However, due to the structure of time series analysis, these approaches tend to yield low clustering performance[3, 10]. The shape centroid is calculated based on the k-Shape algorithm [8]. The algorithm employs NCC to optimally match a pair of time series [3, 8]. DTW barycenter averaging (dba) is a global and iterative method that is developed based on dtw. It randomly selects one time-series as the centroid. Then dtw alignment between time series of the cluster is computed and then updates the centroid. This process iteratively performed until reaching convergence or a certain number of iterations [3, 10]. In partition around medoids (PAM) a time series is selected as the mediod, which has the minimum average distance to all other time series of the cluster. Since the mediod is an element of the original data, it sometimes preferred over median and mean [3].

***Predicting clusters’ labels***: Four well known machine learning algorithms of support vector machine (SVM), logistic regression (LR), neural network (NNET), and tree augmented naïve bayes (TAN) are utilized to predict clusters’ label produced through investigating time series of the sensor data. Four subjects’ variables of “gender”, “age”, “weight”, and “height” are investigated for demonstrating predictability of the clusters’ labels. Since the number of subjects are typically limited in such studies, leave-one-out cross validation method is utilized. Performance of the predictions are measured through “accuracy”, “sensitivity”, “specificity”, and “area under the curve (auc)”. Stacking approach is adopted for elevating the performance of prediction. Therefore, probability of predicting each subjects from the machine learning algorithms is imported as input to random forest algorithm.

***Support vector machine (SVM)***: SVM is a regression and classification algorithm. In the classification application, this algorithm investigates thickest possible margin between categories of data. Hereof, subjects’ variables represent dimensions of space for the SVM algorithm. The subjects on the margin’s border are the support vectors of SVM. The subjects are classified based on their relative position to the margin. SVM searches for the maximum margin and support vectors through utilizing quadratic programming. A generic formulation of a SVM algorithm is demonstrated below [11]:

where:

: Lagrange multiplier

cost: penalty

***Logistics regression (LR)***: LR develops association between a binary outcome (two classes of the subjects) and predictors (subjects’ variables). The outcome of regression equation is logit transformation of classes’ probability. A generic formulation of a LR model is [12]:

***Neural network (NNET)***: NNET mimics biological neural network in translating stimulus (subjects’ data) to information (subjects’ class). In this study, subjects’ data perceived activates the first layer of NNET. Then the effects are transferred to the next layers through multiplying the assigned weights. This consecutive process repeats until the last layer then affects the activation function. Activation function specifies the category of the outcome. Number of layers, and activation functions are investigated through heuristic approaches. A generic formulation of an activation function is demonstrated below [13]:

where:

: inputs applied to neuron

weights

: biased

a(.): activation function

***Tree augmented naïve bayes (TAN)***: TAN is subset of Bayesian belief network (BBN). BBN constructs a stochastic dependency graph to represent conditional dependency among the variables. However, in TAN, the target variable is a parent of the predictors but it does not have any parent variable [14]. Bayes rule is utilized for developing the conditional dependency and a heuristic approach is selecting the structure that has the highest probability. The conditional dependency function for each (i,j) pairs is represented below [15]:

***Leave-one-out cross (LOOCV) validation***: LOOCV is particular case of n-fold cross validation [16]. In this study, there are 15 folds which each fold has 1 subject. Models are trained using 14 folds and then tested over the left out fold. This process repeated until all the subjects are predicted as the test fold.

***Stacking***: Stacking adopts a superior model to combine outcomes of different models for improving the predictive performance. The superior model, produces an optimal ensemble of the sub-models [17]. In this study, probability of each subjects’ cluster is predicted through the four machine learning algorithms (SVM, LR, NNET, and TAN). Then a random forest algorithm (super learner) is developed over the produced probabilities through LOOCV method.

***Random forest (RF)***: RF is a classification and regression algorithm. It develops a forest from ensemble of decision trees. For categorizing each subjects to a cluster, it assigns the vector of subjects’ variables into the decision trees of forest. Each decision tree yields a classification that votes for the clusters. The algorithm assigns the subjects to the clusters based on the majority of the votes [18, 19].

# RESULTS & IMPLICATIONS

***Results***

The subjects reported the lowest Borg’s value during the first nine time epochs, which represents 18% of the overall experiment time. Borg Values remained mostly constant during this time frame. Figure 2 demonstrates mean Borg’s ratings with 95 percent confidence interval for all the subjects in the consecutive time points. As can be seen in fig. 2, the variance of the Borg’s value is constant during the experiment. Also, there is almost linear increment in the Borg values as the experiment continues. This implies the fatigue levels are also increasing mostly linearly as time progresses. However, they cross a threshold after 25 time points that demonstrate higher levels of exertion and fatigue.

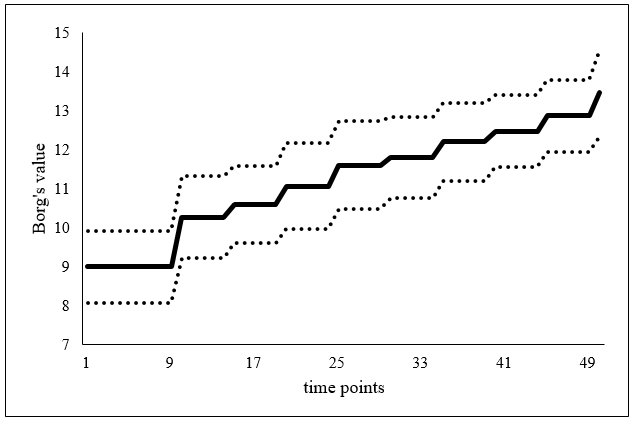
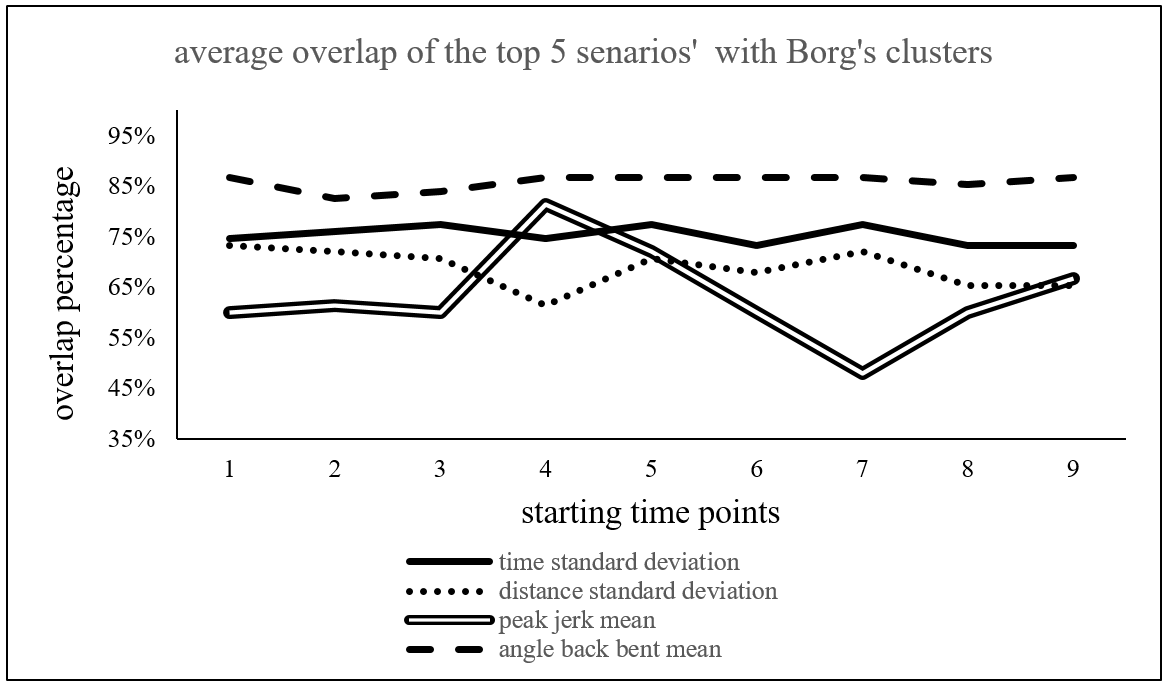


Figure : mean Borg's value for all the subjects

We subsequently used time series clustering by capturing data from 9 different time points, which serves as our first time period in further analysis. Also, there are

Each time series representing a subject is compared with the Borg’s value to understand their similarity (overlap) with the Borg’s value. Figure 3 demonstrates the time series of average Borg’s rate. Each data point represents average of all subjects’ Borg’s rate in a specific time stamp.

In addition to average jerk, peak jerk mean, angle back bent mean, and time standard deviation demonstrated higher accordance with the Borg. However, time standard deviation showed most variability in the deviation. We also observed a more stable pattern in the accordance of all candidate variables with the Borg’s value except for the “peak jerk mean”. There is slightly more accordance with the Borg’s value when earliest time points are considered for time series clustering.



|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| starting time points | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | average |
| time standard deviation | 12.9% | 13.1% | 13.7% | 12.9% | 13.7% | 13.3% | 12.4% | 6.0% | 7.3% | 11.7% |
| distance standard deviation | 7.3% | 7.8% | 5.3% | 2.7% | 9.0% | 9.8% | 7.8% | 5.0% | 2.7% | 6.4% |
| peak jerk mean | 15.2% | 20.0% | 10.3% | 18.1% | 18.1% | 14.0% | 6.5% | 11.9% | 6.0% | 13.3% |
| angle back bent mean | 0.0% | 3.3% | 3.3% | 0.0% | 0.0% | 0.0% | 0.0% | 2.7% | 0.0% | 1.0% |

Figure : Variability of clustering overlap for each candidate variable

As mentioned earlier, partitional time series clustering is a stochastic algorithm and performance of the model is affected by different seeds point as well as centroid and distance measures. The previous table demonstrates variability in clusters’ overlap using top three partitional time series clustering’s setups (centroid and distance measures) for each starting time stamps senatrios. Among the variables, “peak jerk mean” and “time standard deviation” have high variability while “distance standard deviation” and “angle back bent mean” have the highest variability. Although an extensive data is gathered for each subject, an increase in the population test data sets will likely help with better understanding the cause for high variability among some of the chosen clustering outcomes.

These results demonstrate that clustering with sensor data could yield high overlap with Borg’s value with low variability. Also, there is no evidence if starting from which starting time stamp might yield a higher accordance with the actual fatigue “Borg’s value”. The best three combinations of the clusters (starting time point is the first collected data) are selected for voting if a subject belong to a cluster. These three combinations yielded the highest accordance between peak jerk value and Borg’s value. These combinations classified the subjects into two groups. Then predictability of the subjects is evaluated by using subjects’ variables. Four well known machine learning algorithms of SVM, LR, NNET, and TAN are utilized to predict clusters label. The following table demonstrates average performance of all the mentioned machine learning algorithms.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | accuracy | sensitivity | specificity | auc |
| time standard deviation | 56.7% | 0.0% | 77.3% | 86.4% |
| distance standard deviation | 31.7% | 43.8% | 17.9% | 80.8% |
| peak jerk mean | 53.3% | 5.0% | 77.5% | 77.0% |
| angle back bent mean | 35.0% | 21.4% | 46.9% | 71.4% |

Figure 6: Average performance of machine learning algorithms in predicting clusters' labels

The performance of algorithms is fairly low. For improving the predictive performance, the outcome of the models is fed into RF model using the stacking approach.

LOOCV is adopted and each time 14 subjects are considered as training set to predict another subjects. The RF model could predict labels of each subject, correctly. Therefore, 100 percent accuracy achieved.

**Discussion**

This study introduces an approach for understanding employees’ tolerance based on their fatigue levels during the work time. The intuitive approach is interrupting employees’ works and record their fatigue based on Borg’s rates. Then categorizing the employees based their fatigue tolerance through time series clustering through the Borg’s rates. However, practicality of this approach in some working processes that requires private space and high concentration for the employees, is questionable. This study utilized four types of censor data (time standard deviation, distance standard deviation, peak jerk mean, and angle back bent mean) and showed clusters produced by them have fairly high accordance with the cluster made by Borg’s rate. Also, machine learning algorithms could be utilized to predict the clusters developed by the censors’ data. The clusters’ labels were predicted correctly by employing machine learning algorithms over the physiological variables. It means that clusters developed through time series clustering are highly predictable. Therefore, time series clustering and cluster prediction could be utilized together to validate the outcomes.

**Refrences:**

1. Borg, G.A., *Psychophysical bases of perceived exertion.* Med sci sports exerc, 1982. **14**(5): p. 377-381.

2. Hastie, T., R. Tibshirani, and J. Friedman, *The elements of statistical learning New York.* NY: Springer, 2009.

3. Sardá-Espinosa, A., *Comparing time-series clustering algorithms in r using the dtwclust package.* R package vignette, 2017. **12**.

4. Lemire, D., *Faster retrieval with a two-pass dynamic-time-warping lower bound.* Pattern recognition, 2009. **42**(9): p. 2169-2180.

5. Keogh, E. and C.A. Ratanamahatana, *Exact indexing of dynamic time warping.* Knowledge and information systems, 2005. **7**(3): p. 358-386.

6. Berndt, D.J. and J. Clifford. *Using dynamic time warping to find patterns in time series*. in *KDD workshop*. 1994. Seattle, WA.

7. Ratanamahatana, C.A. and E. Keogh. *Everything you know about dynamic time warping is wrong*. in *Third workshop on mining temporal and sequential data*. 2004. Citeseer.

8. Paparrizos, J. and L. Gravano. *k-shape: Efficient and accurate clustering of time series*. in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. 2015. ACM.

9. Cuturi, M. and M. Blondel, *Soft-DTW: a differentiable loss function for time-series.* arXiv preprint arXiv:1703.01541, 2017.

10. Petitjean, F., A. Ketterlin, and P. Gançarski, *A global averaging method for dynamic time warping, with applications to clustering.* Pattern Recognition, 2011. **44**(3): p. 678-693.

11. Kalantar, B., et al., *Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN).* Geomatics, Natural Hazards and Risk, 2018. **9**(1): p. 49-69.

12. Kleinbaum, D.G. and M. Klein, *Logistic regression: a self-learning text*. 2010: Springer Science & Business Media.

13. Karabatak, M. and M.C. Ince, *An expert system for detection of breast cancer based on association rules and neural network.* Expert systems with Applications, 2009. **36**(2): p. 3465-3469.

14. Dag, A., *A Data Driven Framework to Identify the Critical Variables, Visualize Their Conditional Relations and Predict the Outcomes of US Heart Transplants.* 2016.

15. Dag, A., et al., *A probabilistic data-driven framework for scoring the preoperative recipient-donor heart transplant survival.* Decision Support Systems, 2016. **86**: p. 1-12.

16. Stone, M., *Cross-validatory choice and assessment of statistical predictions.* Journal of the royal statistical society. Series B (Methodological), 1974: p. 111-147.

17. Van der Laan, M.J., E.C. Polley, and A.E. Hubbard, *Super learner.* Statistical applications in genetics and molecular biology, 2007. **6**(1).

18. Archer, K.J. and R.V. Kimes, *Empirical characterization of random forest variable importance measures.* Computational Statistics & Data Analysis, 2008. **52**(4): p. 2249-2260.

19. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.