2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Using Machine Learning to Predict Frailty from Cognitive Assessments

Shubham Kumar[1], Chen Du, Sarah Graham, and Truong Nguyen

Abstract— **This paper explores the relation between cognitive and physical aspects of the human body from a machine learning standpoint. We propose to use performance on cognitive assessments to predict frailty of elderly adults with different regression and classification models. We propose a preprocessing scheme with oversampling and imputation to overcome the challenge of an imbalanced data distribution on the existing dataset. We validate the capability of classification models to predict frailty on patients given cognitive input data and provide evidence that machine learning models depend on clinically-defined thresholds.**

I. INTRODUCTION

The human body is a highly interconnected system, in which one can expect intricate dependencies between various domains. Extensive works have been done to establish these connections. A decline in cognitive performance has been associated with a decrease in performance on physical tasks [1]. On the other hand, physical exercise has been established as an effective means to combat cognitive impairment, especially in older adults [2]. This paper explores the relationship between the cognitive and physical functions of the human body from a machine learning standpoint. Specifically, this paper focuses on how machine learning techniques can be used to predict frailty, occurring when deterioration in many systems leads to a heightened susceptibility to stress [2], using information from cognitive assessments.

A model that predicts frailty given cognitive input data allows us to exploit the relationship between cognitive and physical domains of human functioning to make preliminary diagnoses on frailty in patients. In the age of a pandemic, it becomes increasingly imperative to develop options to remotely diagnose people. Physical tests often need safety protocols, supervision, and an in-person proctor to obtain meaningful results (especially for the elderly), but cognitive assessments may be more easily completed remotely with guidance over the phone. The proposed model can pave the way for easier and faster diagnoses, allowing for preventative measures to proactively be taken. Such a system can also prove useful in a future with no ongoing pandemic.

We divide this paper into two main sections. The first explores techniques that can be used to navigate a limited and imbalanced (one in which there is unequal representation of classes) dataset, as is often the case in clinical datasets; the second seeks to demonstrate which cognitive features are most successful in predicting physical function.

Recent work has been done in using artificial intelligence (AI) methods to predict frailty in elderly people. The work in [3] implements different binary classification models to predict physical conditions in a dataset of around one million samples. The study had 58 input variables, ranging from sociodemographic (age and gender) to medical (number of medications) to illnesses (types of diseases). Similarly, [4] predicts frailty in a dataset of 592 patients using up to 70 input variables. These studies use large datasets and many input features to predict physical functioning, but there has been limited work on exploiting the interconnected nature of the cognitive and physical domains from a machine learning standpoint.

II. DATASET & FEATURES

In this section, we will examine the imbalance in the available dataset and discuss motivations for choosing specific features.

The dataset was collected by Jeste et al. [5] and sponsored by International Business Machines Corporation (IBM). It consists of 104 participants from a continuing care senior housing community. Participants consist of 67% female and 33% male with the age ranging between 65-95 years. For this cohort, various sociodemographic, cognitive, physical, and mental variables were collected, and [5] has shown a relation between cognitive function and physical performance. Of the 104 total participants, we include 92 participants for whom we have available data on the variables of interest.

While [5] collected many features, we only consider a subset of those features. First, sociodemographic features include age, gender, and education, which were chosen to provide some background information about each patient to the machine learning model. Second, physical features encompass the Short Physical Performance Battery (SPPB) [6], the Timed Up and Go (TUG) test [7], the physical component of the Medical Outcomes Study 36-item Short Form (SF-36) [9], BMI (kg/m^2), and waist-to-hip ratio (WHR). The TUG and SPPB, in particular, have been identified as effective methods of quantifying function of the lower body [8], while the other features were chosen as supplements.

Lastly, the cognitive features are comprised of the Montreal Cognitive Assessment (MoCA) [10], the University of California San Diego Performance Based Skills Assessment-Brief (UPSA-B) [11], the mental component of the SF-36 [9], and executive function [12]. The MoCA and UPSA-B are popular metrics used to assess overall cognitive and everyday function [13][14], and the others are used as additional features.

The dataset essentially consists of 92 datapoints, making it small for machine learning models such as logistic regression and linear regression. Another challenge we face is the

* Electrical and Computer Engineering & Psychiatry Dept., UC San Diego
[1] s2kumar@ucsd.edu

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

imbalanced distribution among categories. Following the clinically defined cutoffs for frailty in TUG [15][16] and SPPB [17][18] (Table I), SPPB is divided into Frail (0-6) and Pre-Frail+Not-Frail (7-12) for binary classification. The distribution of different frailty level shows that the dataset is significantly imbalanced as only 13.95% and 26.67% of data points are within the frail class for TUG and SPPB, respectively (Table I). An imbalanced dataset makes it difficult for many models to accurately perform classification since these models assume an equal class distribution.

In addition to the imbalance, there are 12 patients (13% of the dataset) with missing values on the MoCA, BMI, and UPSA input features. The missing values further increase the difficulty in using the full dataset to predict frailty.

With the limitations above, it is challenging to efficiently predict physical performance on the TUG and SPPB. To overcome these challenges, we propose a data preprocessing method using imputation and oversampling.

III. DATA PREPROCESSING

First, we observe the effects of imputation on our dataset. Imputation is a method which uses a regressor and the provided input features to estimate any missing features in the dataset. By estimating these features, we can include the patients when training our machine learning model, allowing us to increase the size of the dataset.

We impute the dataset using Bayesian Ridge [19], Decision Tree [20], Miss Forest [21], and KNN Regression [22] as well as basic mean imputation. We implement a logistic regression (LR) classifier trained and evaluated for binary classification on the newly imputed dataset using 5-fold cross validation for a 4:1 train-test split. To reduce the selection bias we repeat 100 trials of randomized cross-validation, and report the average balanced accuracy (BA) for both SPPB & TUG in Table II. BA takes the proportion of each class into account, allowing for a more accurate representation of the classifier's performance on the imbalanced dataset [23].

From Table II, we see that imputation does not have a significant effect on BA for TUG & SPPB, and some of the sophisticated imputation methods (like Miss Forest and KNN) only outperform simply mean imputation by a small margin. This suggests that imputation does not increase the accuracy of TUG and SPPB predictions in the dataset but allows us to expand our dataset.

Second, we implement imbalanced-learn's SMOTE oversampling [24][25]. Oversampling is a technique synthetically creates minority datapoints to balance the dataset and increase the number of datapoints available for training and testing. We synthetically create datapoints up to a ratio (R), where R indicates the ratio of minority to majority datapoints. When R=1, the dataset is perfectly balanced. In Figure 1, we report the average BA of our LR classifier for every 0.1 increment of R over 100 trials and compare it to the BA on the baseline dataset, which is created by deleting patients with missing data. There is no imputation in this process.

TABLE I

CUTOFFS FOR TUG & SPPB

Test	Segmentation	Frail	Pre-Frail	Not-Frail (%)
TUG	Range (seconds)	≥ 14	-	< 14
	Ratio (%)	13.95	-	86.05
SPPB	Range (score)	$0-6$	$7-9$	$10-12$
	Ratio (%)	26.67		73.33

TABLE II

COMPARISON OF DIFFERENT IMPUTATIONS OF SPPB & TUG

Imputer	SPPB BA	TUG BA
Bayesian Ridge	.677	.724
Decision Tree	.668	.729
Miss Forest	.675	.728
KNN	.674	.723
Mean	.672	.715

Figure 1 shows that the BA of both TUG and SPPB increases by oversampling. When R=1, TUG outperforms the baseline by close to 21%, and SPPB improves by 5%. The significant improvement in TUG may come from the fact that the TUG dataset being more imbalanced compared to SPPB (Table I). The TUG dataset has the largest initial gains from oversampling, and then, the effects of oversampling taper off. These results shows that oversampling is an effective method in our dataset to increase samples in the minority class, which effectively balances and increases the size of our dataset.

Lastly, we extract the feature weights to investigate which features are most important for which tests. The feature weights are extracted by gathering the correlation coefficients from the LR classifier. Data preprocessing steps include imputation and oversampling as described above. This process runs for 50 trials, and the normalized and averaged feature weights are reported for both SPPB and TUG in Table III.

Table III shows that age is an important feature in both tests, likely due to the natural decline of the body and

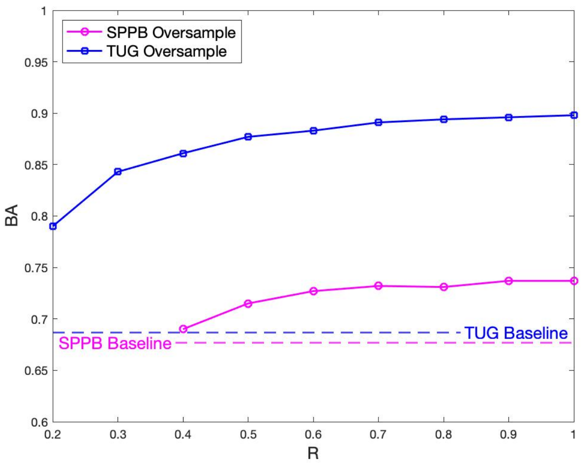

Fig. 1. Effects of Oversampling of TUG & SPPB

TABLE III

FEATURE WEIGHTS FROM LR OF SPPB & TUG

Feature	SPPB Weight	TUG Weight
Age	-0.904	-1.704
Gender	-0.226	-0.114
Education	0.502	-0.448
P-SF-36	0.869	1.01
BMI	-1.064	0.56
WHR	-0.024	0.012
M-SF-36	0.847	1.065
Executive Function	0.312	3.843
UPSA-B	0.049	-1.94
MOCA	0.185	2.288

mind as one ages [26]. Gender, education and WHR have little impact on both tests. We can also see that compared to SPPB, TUG generally has stronger correlations to input features, contributing to its consistently better performance over SPPB. Specifically, TUG has much stronger correlations to cognitive features. In the next section, we give an in-depth treatment to the connection between cognitive and physical function.

IV. EXPERIMENT & RESULT

This section explores the relationship between cognitive features and physical features through the lens of an ablation study. Specifically, we investigate the relationship between cognitive function assessed by the MoCA, UPSA-B, mental component of the SF-36 (M-SF-36), and executive function with physical function determined by performance on the TUG and SPPB. In addition to the given cognitive features, supplementary sociodemographic information is provided as input to the model for every permutation.

A. Method

The primary motivation for choosing the format of an ablation study was to determine which combination of input cognitive features were the best predictors of physical function. This also lets us isolate the distinct impacts of each cognitive feature.

To assess the best machine learning framework from which to understand this relationship given the constraining factors of this dataset, we approach this question from both a classification and a regression standpoint. In our case, we want to understand which approach can be used to derive meaningful results given the size and imbalance of our dataset. For classification, we implement weighted logistic regression (LR). Regression models include linear regression, ridge regression, and SVR. All models are implemented using Python's scikit-learn.

B. Experiment & Results

We exclude the preprocessing steps for the ablation studies because we do not want to introduce synthetic data in the experiment that may bias our result. Subjects with missing features are excluded from experiments. Features are iteratively chosen and then used to train and evaluate the given model on the randomly shuffled dataset on a 4:1 train-test split. Since the dataset is small, we train and evaluate the model for 2000 trials and report the average result, which is more basic application of k-fold cross validation. The central incentive for doing this is to minimize variance across results and report a stable number.

For classification, we use BA to evaluate performance. Regression is evaluated on MAE for both tests, and we round the predictions to the closest whole number to calculate classification accuracy for SPPB. This way, we can directly compare performance of our regression models with our classification models. Note that this process cannot be done for TUG since it is a timed test.

First, we present comparisons between binary classification and regression. Second, we perform arbitrary manipulations of the TUG and SPPB test cutoffs to artificially create a balanced dataset and observe the effects on performance.

1) Classification vs. Regression: We report the top five performing permutations of models and features for TUG and SPPB independently in Table IV. The result indicates that the binary classification model can achieve a relatively high accuracy predicting physical functionality from cognitive features. The best performing features are consistent with the feature weights in Table III, where executive function is the most influential variable for TUG. Interestingly, TUG prediction also performs better than SPPB prediction, as is consistent with the results from data preprocessing in the previous section. We conjecture that this behavior is due to TUG's higher relation to cognitive features (Table III), suggesting that for this dataset, TUG is more highly related to cognitive functioning than SPPB.

We observe from the SPPB results that the classification accuracy for regression is significantly lower than that of binary classification. Furthermore, when we analyze the MAE, the best MAE for TUG is just under 2 seconds, indicating that there is approximately 15% error (the range of TUG scores is 13). For SPPB, the best MAE is nearly 1.9, meaning that there is roughly 14.6% error (the range of SPPB scores is 13). Regression does not provide a reliable prediction of physical function scores. This may be explained by the fact that regression is a finer prediction when compared to binary classification, and achieving robustness in regression requires a much larger, balanced dataset than what we have available. Looking towards binary classification, we are able to obtain some meaningful results and can establish the relationship of cognitive and physical domains of the human body from a machine learning standpoint.

2) Arbitrary Cutoffs: We perform arbitrary manipulations of the clinically defined cutoffs for the SPPB and TUG tests to forcibly create a more balanced dataset for binary classification and discern the following impacts on performance.

As shown in Table I, there is significant imbalance in the dataset when applying the clinically determined cutoffs for frailty for the SPPB and TUG tests. By shifting the cutoffs, we can contrive a more balanced dataset and determine if such a dataset can improve performance, even if results from the resulting model do not have any clinical significance. We experiment with three different cutoffs, each with different

TABLE IV

Cognitive Features for Predicting TUG & SPPB where 1 - M-SF-36, 2 - UPSA-B, 3 - MoCA, 4 - Executive Function

	Binary Classification			Regression			
Test	Model	Feature(s)	BA	Model	Feature(s)	MAE	Acc. (%)
TUG	LR	4	.825	Linear	4	1.956	–
		2,4	.819		1,4	1.963	
		3,4	.812	Ridge	1,4	1.963	
		2,3,4	.776		1,3,4	1.995	
		1,2,4	.752	Linear	2, 4	2.000	
SPPB	LR	4	.614	Linear	1,4	1.875	16.84
		2	.613		1	1.879	20.37
		2,4	.606	Ridge	1	1.879	20.28
		1	.597	Linear	1,2	1.895	20.19
		3,4	.595	Ridge	1,2	1.898	20.10

degrees of balance or imbalance, shown in Table V. There are two important things to note. First, Cutoff #2 of both TUG and SPPB are balanced datasets. Second, SPPB Cutoff #1 is a clinically significant cutoff, created by combining Frail+Pre-Frail (0-9) and keeping Not-Frail (10-12) separate (Table I).

After repeating the process performed in the ablation study described above, we can compare the results from the different manipulations of the dataset to the original, clinically significant division of the dataset.

First, we examine the results for TUG in Table VI and compare it to the original results from Table IV. We observe that for Cutoff #1 and #2, the BA significantly decreases. For Cutoff #3, we observe similar performance when compared to the original results. This result suggests that artificially forcing a balanced dataset by manipulating the cutoffs does not improve classification accuracy. Since the distribution of Cutoff #3 is essentially the inverse of the original distribution, we observe similar results.

TABLE V

Distribution of Arbitrary Cutoffs for TUG & SPPB

Test	Cutoffs #	Frail (%)	Not Frail (%)
TUG	1	27.9	72.1
	2	53.5	46.5
	3	83.7	16.3
SPPB	1	64.4	35.6
	2	53.3	46.7
	3	38.9	61.1

TABLE VI

Prediction on Different Distributions of TUG & SPPB where 1 - M-SF-36, 2 - UPSA-B, 3 - MoCA, 4 - Executive Function

Test	Features	BA Cutoff #1	BA Cutoff #2	BA Cutoff #3
TUG	4	.760	.675	.817
	2,4	.75	.654	.820
	3,4	.744	.666	.811
	1,4	.743	.634	.757
	2,3,4	.735	.645	.783
SPPB	4	.673	.643	.619
	2,4	.663	.628	.612
	3,4	.657	.628	.611
	2	.642	.607	.583
	2,3,4	.641	.617	.612

From the SPPB results in Table VI, we see that the clinically significant Cutoff #1 outperforms the original result. Cutoff #1 has a more balanced dataset compared to the original distribution while remaining clinically significant, which may be why it was able to outperform the original distribution. Cutoff #2 and #3 perform worse than Cutoff #1, but better than the original results.

Even with an unbalanced distribution, the clinically meaningful thresholds yielded better performance than the more balanced datasets with arbitrary cutoffs. This confirms, from a machine learning application, that there is some hidden meaning in the clincal thresholds that holds value to the model. These clincally significant thresholds are able to overcome severe imbalance in the dataset and outperform an arbitrarily created balanced dataset.

V. Conclusion

In this study, we have validated the ability for a machine learning model to classify frailty on the dataset from Jeste et al. [5] using cognitive and sociodemographic input features. Our experiment results indicate that there is a higher cognitive relation with TUG, and have proved that adhering to clinically determined thresholds hold greater innate value to classification models than contrived datasets from an arbitrary cutoff. Lastly, we found that it may be easier for models to classify patients as being frail/at risk of frailty versus being functional. The proposed data preprocessing using SMOTE oversampling increased the BA on TUG and SPPB and allowed us to overcome the inherent imbalance in the dataset.

Future work may look into gathering a larger dataset to explore the scope of robust predictions, completing a similar process the mental domain of the human body, and exploring the possibility of preemptively predicting significant physical decline.

References

[1] M. Tabbarah, E. M. Crimmins, T. E. Seeman, "The Relationship Between Cognitive and Physical Performance: MacArthur Studies of Successful Aging," *The Journals of Gerontology: Series A*, vol. 57, no. 4, pp.228-235, Apr. 2002.

[2] L. Bherer, K. I. Erickson, T. Liu-Ambrose, "A Review of the Effects of Physical Activity and Exercise on Cognitive and Brain Functions in Older Adults," *Journal of Aging Research*, vol. 2013, Sept. 2013.

[3] A. Tarekegn, F. Ricceri, G. Costa, E. Ferracin, M. Giacobini, “Predictive Modeling for Frailty Conditions in Elderly People: Machine Learning Approaches,” *JMIR Medical Informatics*, vol. 8, no. 6, June 2002.

[4] R. C. Ambagtsheer, N. Shafiabady, E. Dent, C. Seiboth, J. Beilby, “The application of artificial intelligence (AI) techniques to identify frailty within a residential aged care administrative data set,” *International Journal of Medical Informatics*, vol. 136, April 2020.

[5] D. V. Jeste et al, “Study of Independent Living Residents of a Continuing Care Senior Housing Community: Sociodemographic and Clinical Associations of Cognitive, Physical, and Mental Health,” *The American Journal of Geriatric Psychiatry*, vol. 27, no. 9, pp. 895-907, Apr. 2019.

[6] J. M. Guralnik et al, “A Short Physical Performance Battery Assessing Lower Extremity Function: Association With Self-Reported Disability and Prediction of Mortality and Nursing Home Admission,” *Journal of Gerontology*, vol. 49, no. 2, pp. M85-M94, March 1994.

[7] D. Podsiadlo, S. Richardson, “The Timed “Up & Go”: A Test of Basic Functional Mobility for Frail Elderly Persons,” *Journal of the American Geriatrics Society*, vol. 39, no. 2, pp. 142-148, Feb. 1991.

[8] J. M. Guralnik et al, “Lower extremity function and subsequent disability: consistency across studies, predictive models, and value of gait speed alone compared with the short physical performance battery,” *The Journals of Gerontology Series A Biological Sciences and Medical Sciences*, vol. 55, no. 4, pp. M221-M231, May 2000.

[9] J. E. Ware, C. D. Shelbourne, “The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection,” *Medical Care*, vol. 30, no. 6, pp. 473-483, June 1992.

[10] Z. S. Nasreddine et al, “The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment,” *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695-699, Apr. 2005.

[11] B. T. Mausbach, P. D. Harvey, S. R. Goldman, D. V. Jeste, T. L. Patterson, “Development of a Brief Scale of Everyday Functioning in Persons with Serious Mental Illness,” *Schizophrenia Bulletin*, vol. 33, no. 6, pp. 1364-1372, Nov. 2007.

[12] D. C. Delis, E. Kaplan, J. H. Kramer, “Delis-Kaplan executive function system,” 2001.

[13] L. Sweet et al, “The Montreal Cognitive Assessment (MoCA) in geriatric rehabilitation: psychometric properties and association with rehabilitation outcomes,” *International Psychogeriatrics*, vol. 23, no. 10, pp. 1582-1591, Aug. 2011.

[14] B. T. Mausbach et al, “Relationship of the Brief UCSD Performance-based Skills Assessment (UPSA-B) to multiple indicators of functioning in people with schizophrenia and bipolar disorder,” *Bipolar Disorders*, vol. 12, no. 1, pp. 44-55, Feb. 2010.

[15] A. Shumway-Cook, S. Brauer, M. Woollacott, “Predicting the Probability for Falls in Community-Dwelling Older Adults Using the Timed Up & Go Test,” *Physical Therapy*, vol. 80, no. 9, pp. 896-903, Sep. 2000.

[16] M. M. Lusardi, G. L. Pellecchia, M. Schulman, “Functional Performance in Community Living Older Adults,” *Journal of Geriatric Physical Therapy*, vol. 26, no. 3, pp. 14-22, Dec. 2003.

[17] J. Subra et al., “The integration of frailty into clinical practice: preliminary results from the Gérontopôle,” *The Journal of Nutrition, Health and Aging*, vol. 16, no. 8, pp. 714-720, Aug 2012.

[18] J. M. Pritchard, “Measuring frailty in clinical practice: a comparison of physical frailty assessment methods in a geriatric out-patient clinic,” *BMC Geriatrics*, vol. 17, no. 1, Nov. 2017.

[19] D. J. C. MacKay, “Bayesian Interpolation,” *Computation and Neural Systems*, Vol. 4, No. 3, 1992.

[20] L. Breiman, J. Friedman, R. Olshen, and C. Stone, “Classification and Regression Trees,” Wadsworth, Belmont, CA, 1984.

[21] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Machine Learning*, vol. 63, no. 1, pp. 3-42, 2006.

[22] N. S. Altman, “An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression,” *The American Statistician*, vol. 46, no. 3, pp. 175-185, Aug. 1992.

[23] K. H. Brodersen, C. S. Ong, K. E. Stephan and J. M. Buhmann, “The Balanced Accuracy and Its Posterior Distribution” in *20th Int. Conf. on Pattern Recognition*, Istanbul, 2010, pp. 3121-3124.

[24] N. V. Chawla, K. W. Bowyer, L. O.Hall, W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, Jun. 2002.

[25] G. Lemaître, F. Nogueira, C. K. Aridas, “Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning,” *The Journal of Machine Learning Research*, vol. 18, no. 1, Jan. 2017.

[26] M. Pinquart, “Correlates of subjective health in older adults: A meta-analysis,” *Psychology and Aging*, vol. 16 no. 3, pp. 414–426, Sep. 2001.