Can we find a connection between physiological indices measured with wearable measuring device?

Yuval Shitrit
Ben Gurion University,
Department of Biomedical Engineering
Beer Sheva, Israel
Yshit@post.bgu.ac.il

Sol Amara
Ben Gurion University,
Department of Biomedical Engineering
Beer Sheva, Israel
Solamar@post.bgu.ac.il

Dan Turetzky
Ben Gurion University,
Department of Biomedical Engineering
Beer Sheva, Israel
dantu@post.bgu.ac.il

***Abstract*— Healthy lifestyle is trending today. Diets and physical training are the basis of healthy lifestyle. Analysis of data about the physiological state of a person and his activity, can be used to improve the efficiency of diet\training programs. For this purpose, we are trying to estimate heart rate (HR) and intensity of activity from data measured with fit-bit wearable device. We found that the most successful model of HR estimation that we created, was achieved by SVM regression for each individual person in the study (highest success rate of 72% for 5BPM error margin). Individuals with more measurements lead to better results. We believe that finding the optimal time frame for data collection will improve the results significantly. As people's normal physiological state changes over time due to many things (e.g., aging, disease, and the results of the dieting and activity as well) we need to update the model occasionally.**

Keywords*— heart rate, intensity, calories, steps, MET, regression.

I. Introduction

Physical activity can greatly affect the physiological condition of a person and vice versa. We want to evaluate such connections, with the hope that these connections can be used to improve health related programs, such as training programs, and diets.

We found our data set in Kaggle. The data contains measurements made with fit-bit devices, which include daily, hourly and per minute measurements of the following – calories expenditure, intensity (four categorizes of intensity of activity classified by fit-bit algorithms), steps, heart rate, MET (metabolic equivalents), sleep, and weight [1]. The data was collected from 30 people for two months. Heart rate is an important physiological measurement that can indicate about serval pathologies such as mental condition, physical endurance, problems in the nervous and hormonal systems, etc. For this reason, we chose to estimate heart rate from information of steps, calories, and intensity. we decided to estimate the heart rate per minute (instead of seconds, which is the unit the data was measured in), because secondly differences of the heart rate don't indicate much about long term physiological processes, affected by physical activities. At a later stage, we chose to also try to estimate intensity levels of activities from data of heart rate, steps, and calories.

As mentioned, we used a database from Kaggle. About 1500 people who used this data base uploaded their work [2]. Their code presents a way to clean, organize the data, present it (e.g., two features plotted against each other with a choice of ID), and analyze it (e.g., correlation matrix of different features). Even so, we didn't find in these works an attempt to predict one parameter from the others, nor did we find conclusions from the data representation\analysis about connections they found between the different parameters. In the literature we haven't found similar research to ours, but there are works about relations between the same measurements that we used (and related ones, such as velocity of movement). In some of those works, the measures were also taken using wearable tracking devices. Between those papers, there is one which presented a successful prediction of heart rate from distance, velocity, and acceleration [3]. Others presented a problem of estimating caloric expenditure from heart rate, as heart rate is affected by non-activity related things such as emotional state and medications [4]. This result might indicate a use for a more complex model. Another research explored the connection between step rate and MET [5]. There are countless more papers about exploration of these connections and related ones [6] [7], and some use machine learning tools for that purpose [8]. From all the above, we conclude that there is a justification for our attempt at creating such models.

II. Methods

First, we cleaned and organized the data, which contained millions of data points. Sleep and weight data were unusable since they were recorded irregularly and on very partial times out of the two months. Then we organized and cleaned the data of calories, steps, heart rate, intensity and MET to measurements per minute. We divided the data to 80% train data and 20% test data. To evaluate heart rate from the other measurements we created a linear regression model. We used MET and intensity separately to create the regression model and they both contributed similarly to the model, so we removed the MET from the model. To check if all remaining features were valuable, we used Ridge and Lasso regressions. We then normalized the heart rate by reducing the average BPM of each person from his measurements which improved the results. Even so, the results of the model were not very promising, so we decided to look for a new approach of using the data to achieve our original goal (building training programs, diets, etc.). We chose two approaches. The first approach was to try a different model for HR estimation. We used SVM regression on the same features. In an attempt to improve the model even further, we used SVM regression on each individual separately. The second was to estimate MET as an intensity measurement. The intensity levels are categorized by Fitbit algorithms to 4 levels – Sedentary, light, moderate and very active. At first, we categorized the MET data to 4 levels with equal bin width to present the intensities, because the intensity data is extremely unbalanced. As this method doesn't necessarily represent the categorization, we created a model using the unbalanced intensity data and compared between the two.

III. Results

In tables 1,2 and 3 we see that the Lasso and Ridge regression didn't improve the results. The resulted thetas (regression coefficients) weren't close to zero, so we can say that all the 3 features (calories, steps and intensity) are valuable. Despite all the features are relevant, we still need to improve the model because the results weren't significant (p-value - 0.333). To do so, we subtracted the mean HR of each individual from his HR data. This linear regression model is:

$$HR = 4.9035 \cdot calories - 0.0052 \cdot steps + 8.0813 \cdot intensity - 12.618$$

The success rate appears in table 4.
p-value – 0.9519
Since these results weren't very good either, we tried using SVM regression - p-value – 0.1485, the success rate appears in table 5. Since SVM regression provided the best results, we built a model for each individual with SVM regression (and normalized BPM). The results appear in table 6.

To explore a different approach to achieve the project's goal, we categorized intensity to four levels. First, we estimated MET as intensity measure from HR, calories, and steps. We used classification tools of Linear SVM, gaussian EM and Fine tree. The algorithm which produced the highest success rate was the Fine tree – 99.48%. Secondly, the algorithm which produced the highest success rate for the intensity was the Gaussian EM – 98.0%. The success rate for the other algorithms appears at tables 7,8. Other results appear in the references.

IV. Discussion, conclusions and future work

When trying to estimate heart rate, the success rate of the linear regression model was under 50%, even when allowing a 5bmp error margin. The Lasso and Ridge regression didn't improve the results and none of the coefficients converged to zero, which indicates all the features used are valuable. Since each person is in a different normal physiological state, we subtracted the average heart rate of each individual in an attempt to improve individual bias, which indeed improved the results (35% to 45% with 5bpm error margin), but the results remained poor. Using a non-linear regression method with SVM regression we achieved better results (48.63% with 5bpm margin error), which indicates the possibility of a non-linear connection between heart rate and the different parameters. To further improve the results, we created a SVM regression model for each individual. From the individuals with high significance (p-value<0.012) the ones with the highest success rate are subjects 7 and 11 (72.23% and 63.31% respectively, for 5bpm error margin). This subjects also have the highest number of measurements, which might indicate that a larger data set (taken for a specific person over a larger amount of time) would improve the results. The concept of improving the results with a larger data set is a basic concept, but we believe there is an important point to address in our case. As a person trains\diets his normal physiological state might change over time and affect for example his average heart rate. For this reason, we believe that there is a specific optimal data set, which balances between a large enough data set and the fact that the physiological state of a person can change over time so a too large of a data set would not consider the time-related physiological state of a person.

When trying to estimate the intensity, we were faced with a problem of unbalance between the different intensity categorizes. As most of the time people are stationary or moving with low effort, this causes a very big unbalance in the categorized data. To avoid this unbalance, we divided the continuous MET data, as an intensity value, into four groups with equal width, which can be seen clearly in fig 3 and fig 4. The results of this partition, using any supervised learning are very successful (>95%). As we can't be sure whether this partition of the data represents the real division for intensities (which are partitioned to categorizes using fit-bit algorithms), we preformed the analysis on the unbalanced intensity data as well. The results were also very high (>95%). From the confusion matrix of both models (fig 5 and fig 6) we can see that the division we created for the continuous data doesn't reflect the real intensities. This is mainly seen in fig 5, by the fact that the miss labels appear only in neighboring categorizes. In the confusion matrix of the unbalanced model (fig 6), we can see more spread of the values, which would suit a real partition of data to categorizes.

To conclude, the heart rate estimation of each individual with SVM regression, and the gaussian EM model for intensity categorization produced models that might be good enough to be used for the purposes of our project. We believe there is a lot more that can be done. As seen, there is research about connection between many of the parameters we checked and others that are related. Using this research and by gathering new data as well, we can create new models or implement changes to our models. Adding relevant features (such as velocity), that weren't measured in our data set, might improve the models. As mentioned, we believe that there is an optimal data set length for each individual (which can also change over time). We can approach this concept in different ways. One way to approach it, is to study how different people with different time frames affect the result, and to find an optimal time frame in the sense of one that is generally good for many people. A different approach is to use the optimization algorithm on an individual until it achieves certain demands (e.g., certain success rate). For both approaches, we would need to update the algorithm occasionally. We believe the first one would require a less frequent update as it generalizes for the general population, and the second approach would require more frequent updates as there is more bias involved. A way to deal with the unbalance of the data when estimating intensity or MET, is to take the measurements only during waking hours because that's a lot of inactive hours that make the data unbalanced.

Acknowledgment

Furberg, Robert; Brinton, Julia; Keating, Michael ; Ortiz, Alexa

The data used in this work was taken from Kaggle. These datasets were generated by respondents to a distributed survey via Amazon Mechanical Turk between 03.12.2016-05.12.2016. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. Individual reports can be parsed by export session ID (column A) or timestamp (column B). Variation between output represents use of different types of Fitbit trackers and individual tracking behaviors / preferences. [9]

REFERENCES

[1] "Fitabase Data Dictionary." https://www.fitabase.com/resources/knowledge-base/exporting-data/data-dictionaries/ (accessed Aug. 02, 2022).

[2] "Fitbit EDA." https://kaggle.com/code/rakgyunim/fitbit-eda (accessed Aug. 05, 2022).

[3] S. Maenaka, H. Suwa, Y. Arakawa, and K. Yasumoto, "Heart Rate Prediction for Easy Walking Route Planning," *SICE Journal of Control, Measurement, and System Integration*, vol. 11, no. 4, pp. 284–291, Jul. 2018, doi: 10.9746/jcmsi.11.284.

[4] J. M. Janot, "Calculating caloric expenditure: optimize clients' workouts by using the ACSM metabolic equations to determine exercise intensity and caloric expenditure," *Idea fitness journal*, vol. 2, no. 6, pp. 32-, 2005.

[5] M. Abel, J. Hannon, D. Mullineaux, and A. Beighle, "Determination of step rate thresholds corresponding to physical activity intensity classifications in adults," *J Phys Act Health*, vol. 8, no. 1, pp. 45–51, Jan. 2011, doi: 10.1123/jpah.8.1.45.

[6] L. H. Epstein, R. R. Wing, and J. K. Thompson, "The relationship between exercise intensity, caloric intake, and weight," *Addictive Behaviors*, vol. 3, no. 3, pp. 185–190, Jan. 1978, doi: 10.1016/0306-4603(78)90017-5.

[7] J. MAYER, P. ROY, and K. P. MITRA, "Relation between Caloric Intake, Body Weight, and Physical Work: STUDIES IN AN INDUSTRIAL MALE POPULATION IN WEST BENGAL," *The American Journal of Clinical Nutrition*, vol. 4, no. 2, pp. 169–175, Mar. 1956, doi: 10.1093/ajcn/4.2.169.

[8] M. Nipas, A. G. Acoba, J. N. Mindoro, M. A. F. Malbog, J. A. B. Susa, and J. S. Gulmatico, "Burned Calories Prediction using Supervised Machine Learning: Regression Algorithm," in *2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T)*, Mar. 2022, pp. 1–4. doi: 10.1109/ICPC2T53885.2022.9776710.

[9] R. Furberg, J. Brinton, M. Keating, and A. Ortiz, "Crowd-sourced Fitbit datasets 03.12.2016-05.12.2016." Zenodo, 2016. doi: 10.5281/zenodo.53894.

TABLE I. LINEAR REGRESSION RESULT

$error$	Success rate	
	$threshold$ 2 bpm	$threshold$ 5 bpm
8.8129	14.2262%	34.9415%

TABLE II. LASSO RESULT

λ	$error$	Success rate	
		$threshold$ 2 bpm	$threshold$ 5 bpm
10^{-3}	12.4825	9.1492%	23.1554%
10^{-4}	10.3581	11.4215%	28.2246%
10^{-6}	8.8462	13.9369%	34.4785%
10^{-7}	8.8462	13.9369%	34.4785%

TABLE III. RIDGE RESULT

λ	$error$	Success rate	
		$threshold$ 2 bpm	$threshold$ 5 bpm
10^{-3}	11.0214	10.7062%	26.5985%
10^{-4}	10.2793	11.7431%	29.1523%
10^{-6}	9.1127	13.5431%	33.5292%
10^{-7}	8.9061	14.3631%	34.8508%

TABLE IV. NORMALIZED BPM RESULTS

$error$	Success rate	
	$threshold$ 2 bpm	$threshold$ 5 bpm
6.9163	19.3492%	45.2169%

TABLE V. NORMALIZED BPM WITH SVM RESULTS

$error$	Success rate	
	$threshold$ 2 bpm	$threshold$ 5 bpm
0.0761	21.083%	48.63%

TABLE VI. NORMALIZED BPM WITH SVM RESULT- MODEL FOR EACH SUBJECT SEPARATALY

$subjec$	$num\ of\ meas$	$p-value$	$error$	Success rate	
				$threshold$ 2 bpm	$threshold$ 5 bpm
1	20624	0.135	0.197	23.07%	53.03%
2	401	0.999	-2.96	15.94%	39.38%
3	22070	$1.85 \cdot 10^{-5}$	0.527	25.32%	58.19%
4	10929	0.161	0.217	23.34%	53.28%
5	37433	$8.89 \cdot 10^{-9}$	0.555	26.27%	58.76%
6	25885	0.191	0.0936	23.62%	53.45%
7	39360	$5.49 \cdot 10^{-11}$	0.477	34.69%	72.23%
8	34088	0.188	0.138	31.28%	65.82%
9	24265	0.012	0.286	20.98%	49.20%
10	4885	0.141	0.396	24.67%	54.55%
11	40753	$2.03 \cdot 10^{-4}$	0.344	28.68%	63.31%
12	19581	0.974	-0.272	26.31%	58.25%
13	19550	0.851	-0.137	22.34%	52.26%
14	26795	0.093	0.264	24.85%	55.82%

TABLE VII. CLASSIFICATION- MET RESULTS

Model type	$Linear\ SVM$	$Gaussian\ EM$	$Fine\ tree$
Success rate	98.79%	95.28%	99.48%

TABLE VIII. CLASSIFICATION- INTENSITY RESULTS

Model type	$Linear\ SVM$	$Gaussian\ EM$	$Fine\ tree$
Success rate	97.4%	98.0%	96.9%

Fig. 1. Example of a tetha's convergence from linear regression

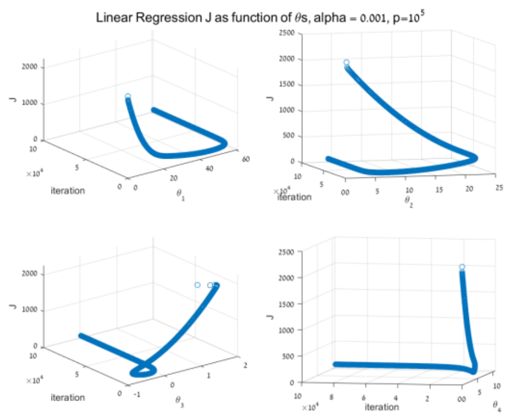

Fig. 2. Example of HR prediction and the true values for subject 7, when the black line present 5BPM margin

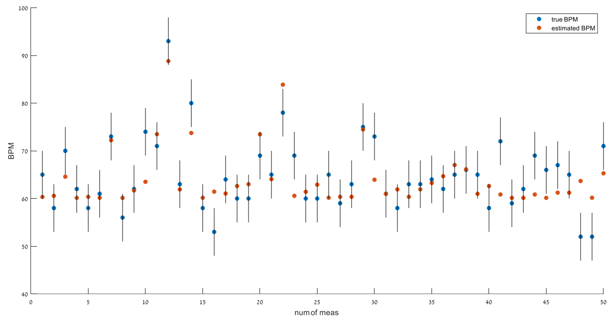

Fig. 3. Example of linear SVM categorization for the MET

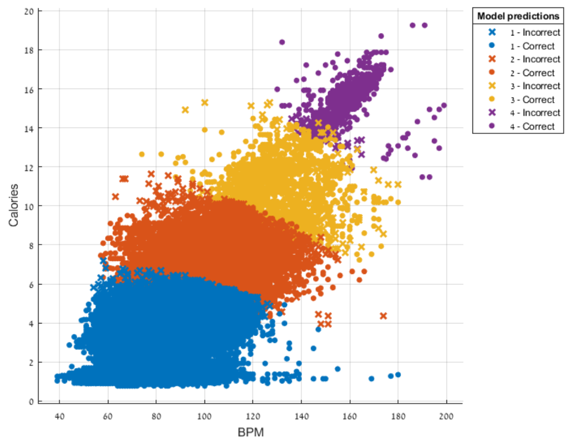

Fig. 4. Example of linear SVM categorization for the intensity

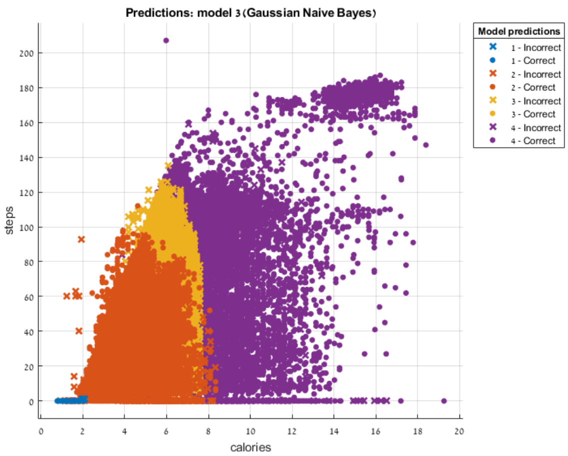

Fig. 5. Example of the confusion matrix of fine tree categorization for the MET

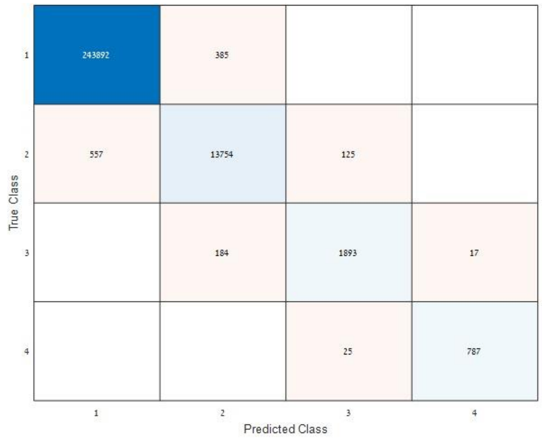

Fig. 6. Example of the confusion matrix of gaussian EM categorization for the intensity