# Evaluation of Machine Learning Algorithms to Detect Irregular Health States in Wearable Sensor Generated Data

Reto Wettstein[1,2]
[1]Department of Medical Information Systems
University Hospital Heidelberg
Heidelberg, Germany
e-mail: reto.wettstein@med.uni-heidelberg.de

Christian Fegeler[2]
[2]Department of Medical Informatics
Heilbronn University of Applied Sciences
Heilbronn, Germany
e-mail: christian.fegeler@hs-heilbronn.de

***Abstract*—Wearable devices facilitate continuous monitoring of personal health data. However, automated health state analysis based on this data is challenging in various aspects. This work presents preliminary algorithm evaluation results for health state irregularity detection based on a continuous data sample collected by an in-ear heart rate and body temperature sensor. The results show that a One-Class Support Vector Machine could be suitable for the task.**

***Keywords–Algorithm Evaluation; Anomaly Detection; Health State; Wearable Generated Data.***

## I. Introduction

Mobile devices like smart watches and fitness trackers are becoming an integral part of our lives. This facilitates continuous monitoring and analysis of personal health data outside of clinical environments [1]. There are already applications which use the ability of wearable devices for specific disease monitoring, like the heart arrhythmia detection functionality by the Apple Watch [2] or the Empatica Embrace 2 seizure detection bracelet [3]. But, none of them considers a person's overall health state. Based on this background, we have built a prototype of a real-time monitoring system for automated irregular health state detection [4]. The centerpiece of this system is a machine learning server component, deciding whether measurements are normal or indicate a change in a persons health state.

Interpretation of sensor health data is challenging. Physiological Response Patterns (PRP) depend on many factors like activities, the environmental context or demographic data and can change over time [5]–[7]. Therefore, PRPs can not be described in general terms. Algorithms have to be trained on a person related basis. Additionally, it is often difficult to collect and access irregular PRPs [8] and thus, this data is not available during training of algorithms. Furthermore, not only the accuracy of applied algorithms plays a key role, also sensitivity and specificity need to be taken into account to reduce, for example, alarm fatigue [9].

However, anomaly detection algorithms could be one type of algorithms used to classify individual PRPs as either normal or irregular. They have been successfully used in other domains (e.g., credit card fraud detection or measurement error detection) where irregular data is not available or can change over a period of time [10].

The objective of this work is to evaluate four anomaly detection algorithms in the context of wearable sensor generated health data. The aim is to verify whether anomaly detection algorithms, already successfully used in other fields than medicine, are suitable for the above mentioned system to detect irregularities in continuously measured health data like body temperature and heart rate.

The remainder of this work is structured as follows: Section II describes the approach for data collection, data preparation and evaluation of the selected algorithms. In Section III, the classification results of the algorithms are presented. Finally, a conclusion and an outlook about future work is given in Section IV.

## II. Methods

The anomaly detection algorithms have been selected so that they are based on different mathematical concepts. The selected ones are Local Outlier Factor, Isolation Forest, One-Class Support Vector Machine and Autoencoder. For assessment of these algorithms, a 72 hour-long data sample of a healthy 28 year old male subject (N = 1) was recorded using the prototype. Utilizing an in-ear sensor, the vital signs body temperature and heart rate were measured in 5 second intervals. For later division of the collected data into training and test set, the measurements were labeled according to the performed activity. After collection of the sample, the data was split into 2 minutes long time-series, having an overlap of 30 seconds (N = 6200). Since generation of irregular health data is not possible at the push of a button, the measurements during the activities sport, metro and eating were regarded as artificial irregularities. All the remaining measurements were considered as normal. Training of the algorithms was based on a data-driven approach supplement with two statistical features (i.e., mean and standard deviation of heart rate and body temperature). Only the normal data was used for training. Finally, the anomaly detection algorithms were evaluated on the artificial irregular data using confusion matrices to calculate the metrics accuracy, sensitivity and specificity.

## III. Results

The performance evaluation of the algorithms was done in an overall setting and individually for each type of irregular data. In the overall setting (see Table I), the algorithms performed with an accuracy higher than 80 %. With the exception of the Isolation Forest, specificity was higher than sensitivity. All algorithms showed a specificity higher than 88 %. For sensitivity, the algorithms reached results better than 76 %. The best overall results were achieved using the One-Class Support

TABLE I. Overall results showing the confusion matrix (−1 irregular, +1 normal), accuracy, sensitivity and specificity of each algorithm using all types of irregular data combined.

| | | Local Outlier Factor | | Isolation Forest | | One-Class SVM | | Auto-encoder | |
|---|---|---|---|---|---|---|---|---|---|
| Confusion Matrix | | −1 | +1 | −1 | +1 | −1 | +1 | −1 | +1 |
| | −1 | 142 | 22 | 148 | 16 | 144 | 20 | 125 | 39 |
| | +1 | 14 | 150 | 24 | 140 | 13 | 151 | 19 | 145 |
| Accuracy | | 89.02 % | | 87.80 % | | 89.94 % | | 82.32 % | |
| Sensitivity | | 86.59 % | | 90.24 % | | 87.80 % | | 76.22 % | |
| Specificity | | 91.46 % | | 85.37 % | | 92.07 % | | 88.41 % | |

Vector Machine with 89.94 % accuracy, 87.80 % sensitivity and 92.07 % specificity.

Regarding each type of irregular data individually (see Table II), the activity sport was identified best in all four algorithms. The Local Outlier Factor and the Autoencoder performed better for the type eating than for the type metro. The One-Class Support Vector Machine and the Isolation Forest reached better results for the type metro than for the type eating.

## IV. Conclusion and Future Work

This work shows preliminary results regarding the ability of anomaly detection algorithms to classify PRPs of activities collected by wearable devices as either normal or irregular. The use of these algorithms in combination with the prototype could potentially enable individuals to identify aggravations of their health earlier and thus, seek medical attention earlier.

The advantage of the selected algorithms is that they consider the great inequality in the distribution of the two kinds of data, normal and irregular. By changing the deciding threshold between normal and irregular time-series measurements, it is possible to take influence on the sensitivity and specificity of an algorithm. This is of most importance in medical applications and allows to change the focus between not missing any true positive or not having to many false positive classifications. For specific applications, this tradeoff would have to be individually reviewed. The most promising results for prototype use were shown by the One-Class Support Vector Machine. However, the other algorithms have also shown positive results, so that a majority vote could be considered, if computationally reasonable.

Further research is needed to assess whether the same results can be achieved for more subjects and if irregularities caused by an imminent or already occurring disease could also be detected with high accuracy, sensitivity and specificity. Additionally, adding more monitoring resources for vital signs, such as blood pressure and respiratory rate, could improve the results.

TABLE II. Specific results showing the confusion matrix (−1 irregular, +1 normal), accuracy, sensitivity and specificity for each algorithm using each type of irregular data individually.

| | | | Sport | | Metro | | Eating | |
|---|---|---|---|---|---|---|---|---|
| Local Outlier Factor | Confusion Matrix | | −1 | +1 | −1 | +1 | −1 | +1 |
| | | −1 | 63 | 1 | 37 | 15 | 42 | 6 |
| | | +1 | 6 | 58 | 5 | 47 | 3 | 45 |
| | Accuracy | | 94.53 % | | 80.77 % | | 90.63 % | |
| | Sensitivity | | 98.44 % | | 71.15 % | | 87.50 % | |
| | Specificity | | 90.63 % | | 90.38 % | | 93.75 % | |
| Isolation Forest | Confusion Matrix | | −1 | +1 | −1 | +1 | −1 | +1 |
| | | −1 | 64 | 0 | 51 | 1 | 33 | 15 |
| | | +1 | 9 | 55 | 8 | 44 | 7 | 41 |
| | Accuracy | | 92.97 % | | 91.35 % | | 77.08 % | |
| | Sensitivity | | 100 % | | 98.08 % | | 68.75 % | |
| | Specificity | | 86.15 % | | 84.62 % | | 85.42 % | |
| One-Class SVM | Confusion Matrix | | −1 | +1 | −1 | +1 | −1 | +1 |
| | | −1 | 64 | 0 | 45 | 7 | 35 | 13 |
| | | +1 | 5 | 59 | 5 | 47 | 3 | 45 |
| | Accuracy | | 96.09 % | | 88.46 % | | 83.33 % | |
| | Sensitivity | | 100 % | | 86.54 % | | 72.92 % | |
| | Specificity | | 92.19 % | | 90.38 % | | 93.75 % | |
| Auto-encoder | Confusion Matrix | | −1 | +1 | −1 | +1 | −1 | +1 |
| | | −1 | 64 | 0 | 29 | 23 | 32 | 16 |
| | | +1 | 7 | 57 | 6 | 46 | 6 | 42 |
| | Accuracy | | 94.53 % | | 72.12 % | | 77.08 % | |
| | Sensitivity | | 100 % | | 55.77 % | | 66.67 % | |
| | Specificity | | 89.06 % | | 88.46 % | | 87.50 % | |

## References

[1] M. J. Bietz et al., “Opportunities and challenges in the use of personal health data for health research,” Journal of the American Medical Informatics Association, vol. 23, no. e1, 2016, pp. e42–e48.

[2] Apple Inc., “Using Apple Watch for Arrhythmia Detection,” 2018, URL: https://www.apple.com/healthcare/site/docs/Apple_Watch_Arrhythmia_Detection.pdf [accessed: 2019-01-31].

[3] Empatica Inc., “Embrace 2, Medical quality technology for epilepsy management,” 2018, URL: https://www.empatica.com/en-eu/embrace2/ [accessed: 2019-01-31].

[4] R. Wettstein and C. Fegeler, “Real-Time Body Temperature and Heart Rate Monitoring System for Classification of Physiological Response Patterns using Wearable Sensor and Machine Learning Technology,” Department of Medical Informatics, Heilbronn University of Applied Sciences, unpublished.

[5] C. Bouchard and T. Rankinen, “Individual differences in response to regular physical activity,” Medicine and science in sports and exercise, vol. 33, no. 6 Suppl, 2001, pp. s446–s451; discussion s452–s453.

[6] Z. Obermeyer, J. K. Samra, and S. Mullainathan, “Individual differences in normal body temperature: longitudinal big data analysis of patient records,” BMJ (Clinical research ed.), vol. 359, 2017, p. j5468.

[7] F. Shamout, D. Clifton, and T. Zhu, “Age- and Sex-Based Early Warning Score,” 2017, URL: http://www.robots.ox.ac.uk/~davidc/pubs/transfer_fs.pdf [accessed: 2019-01-31].

[8] C. H. Lee and H.-J. Yoon, “Medical big data: promise and challenges,” Kidney research and clinical practice, vol. 36, no. 1, 2017, pp. 3–11.

[9] S. Sendelbach and M. Funk, “Alarm Fatigue, a patient safety concern,” AACN Advanced Critical Care, vol. 24, no. 4, 2013, pp. 378–386.

[10] H. P. Kriegel, P. Krger, and A. Zimek, “Outlier Detection Techniques,” Tutorial at the 16th ACM International Conference on Knowledge Discovery and Data Mining, 2010, URL: http://www.dbs.ifi.lmu.de/~zimek/publications/KDD2010/kdd10-outlier-tutorial.pdf [accessed: 2019-01-31].