Person identification using deep neural networks on physiological biomarkers during exercise

Wang, Zuowen ; Wang, Shu ; Lafaye, Céline ; Saubade, Mathieu ; Gremeaux, Vincent ; Liu, Shih-Chii

DOI: https://doi.org/10.1109/BioCAS54905.2022.9948570

Person identification using deep neural networks on physiological biomarkers during exercise

Zuowen Wang[1], Shu Wang[1], Céline Lafaye[2], Mathieu Saubade[2], Vincent Gremeaux[2], Shih-Chii Liu[1]
[1]*Institute of Neuroinformatics, University of Zurich and ETH Zurich*
Zurich, Switzerland
[2]*Sports Medicine Unit, Swiss Olympic Medical Center, Division of Physical Medicine and Rehabilitation*,
Lausanne University Hospital, Lausanne, Switzerland
zuowen,shu,shih@ini.uzh.ch, Celine.Lafaye@chuv.ch

***Abstract*—Much progress has been made in wearable sensors that provide real-time continuous physiological data from non-invasive measurements including heart rate and biofluids such as sweat. This information can potentially be used to identify the health condition of a person by applying machine learning algorithms on the physiological measurements. We present a person identification task that uses machine learning algorithms on a set of biomarkers collected from 30 subjects carrying out a cycling experiment. We compared an SVM and a gated recurrent neural network (RNN) for real-time accuracy using different window sizes of the measured data. Results show that using all biomarkers gave the best results from any of the models. With all biomarkers, the gated RNN model achieved ∼90% accuracy even in a 30 s time window; and ∼92.3% accuracy in a 150 s time window. Excluding any of the biomarkers leads to at least 7.4% absolute accuracy drop for the RNN model. The RNN implementation on the Jetson Nano incurs a low latency of ∼45 ms per inference.**

***Index Terms*—machine learning, physiology, wearable device**

I. Introduction

Wearable sensor technology or wearables can provide continuous real-time physiological or brain signal information useful for identifying the health condition of an individual such as for cardiovascular care [1], sleep monitoring [2] and sports physiology [3]. Possible applications in sports physiology with wearable devices include workload and hydration monitoring [3], [4]. There is increasing interest in wearable sweat devices that provide continuous real-time monitoring of sweat biomarkers [5], [6], because biomarkers such as sweat sodium concentration and sweat rate are linked to electrolyte loss and fluid balance [7]. In [8] the authors used sweat and physiological biomarkers for predicting the hydration status of an individual, and demonstrated the potential of providing a personalized hydration strategy using these biomarkers.

Various machine learning (ML) algorithms have been utilized in recent studies and products for analyzing physiological signals. In [9] the authors used a deep denoising convolutional auto-encoder under a self-supervised training scheme to implement an autonomous pain assessment system with electrocardiography (ECG), electromyography (EMG) and electrodermal activity (EDA). In [10] the authors proposed a method for extracting a representation of physiological signals, excluding components of predefined feature engineering. Their method could be used to examine existed scientific hypotheses and potentially discover new findings. A few studies have looked at person identification by applying ML methods, including support vector machine (SVM) and recurrent neural networks (RNN), to ECG [11] [12] signals. This task is potentially useful for setting up individualized profiles.

In this work, we aim to study the potential of using ML methods on real-time biomarker recordings for a person identification task. The recordings come from sensors that measure heart rate, biking power, body temperature and sweat rates on the back and arm of 30 subjects in a cycling exercise study. We want to predict the unique person identification label of each subject with the recordings from different sensors. We investigate different factors that can impact the accuracy of the task: 1) time window sizes of recordings needed to achieve high accuracy; 2) the importance of the different sensor types; 3) the accuracy achieved by different ML methods including RNNs. We also determine the latency of running the RNN on a low-power edge platform. Section II describes the physiological data collection protocol and Section III describes the machine learning methods used in this work and Section IV provides the results.

II. Data collection protocol and biomarkers

In this section, we describe the protocol for collecting the data used in this study, including heart rate (**hr**), core temperature (**ct**), sweat rate measured from the arm (**sra**) and back (**srb**); and the instantaneous power output (**pw**).

A. Subjects

Participants were recruited from triathlon and cycling clubs around Lausanne, Switzerland. Criteria for participation were that the subject should be between 18 and 50 years old, without known medical condition or regular use of any medication. Thirty eligible athletes (26 men and 4 women; age 33±9 yo; weight 69.5±8.2 kg; height 178±7cm; VO_2max 56.1±7.2 ml/kg/min) volunteered for the study and gave their written informed consent to participate. This study was approved by the ethical commission of the Canton of Vaud Switzerland on 10.05.2021 (Protocol No. 2019-01235). Prior to the start of

This work was partially funded by SNSF-Sinergia WeCare project (N°CRSII5_177255).

the study, all participants underwent a medical examination to ensure their health condition complied with the criteria of this study. The participants also completed a graded cycling test (adapted from [13]) to determine the maximal HR (HRmax), maximal aerobic power (MAP), and maximal oxygen uptake (VO_2max highest value of 20 seconds average).

B. Procedure

All sessions took place in a laboratory at the unit of sport medicine center at Lausanne University Hospital (Lausanne, Switzerland) between May and December 2021, and were conducted by the same experimenter.

To ensure similar fitness level and hydration status among subjects, they were asked to comply with the following specifications: 24 hours before the onset of a session, all participants were required to abstain from caffeine, alcohol, or vigorous exercise and to drink a certain quantity of water calculated using the participants' body weight to make sure the participants were well hydrated right before the session started. In addition, around 9 hours before the start of a session, participants were also instructed to ingest an ingestible telemetric temperature system for measuring core temperature during the sessions.

Upon arrival at the laboratory, the resting heart rate, blood pressure, and aural temperature of the participants were measured. If the resting blood pressure and/or the resting aural temperature were high, the session was postponed to another day. Participants were then asked to void the bladder and to provide a urine sample for determining the baseline pre-cycling urine specific gravity (USG, Model UG-1, ATAGO; Bellevue, WA). A USG smaller than 1.020 g/mL indicated a well-hydrated state and was necessary for continuation, otherwise the participants were asked to drink 500 mL of water within 30 mins before the onset of cycling.

Participants then cycled on a stationary ergometer (CycleOps Pro 400, CycleOps, Madison, USA) equipped with power meter pedals in a room with controlled environment. All participants cycled for a total of 100 mins, including 10 mins warm-up (E0) and three main cycling phases (E1, E2, and E3), each lasting 30 mins. The intensity for E1, E2, and E3 were 40%, 70%, and 30% MAP, respectively.

C. Biomarkers

Figure 1 shows example sensor recordings from a subject in an exercise session. Heart rate (**hr**) was recorded using a HR belt (H7, Polar Electro OY, Kempele, Finland) at a sampling rate of 1 Hz throughout E1, E2, and E3. Core body temperature (**ct**) was sampled every 30 s using the e-Celsius™ capsule (BodyCap, Caen, France) for all three main phases.

Sweat rates were measured on two locations (arm and back, **sra** and **srb**) with two Kudusmart® monitors (Crossbridge Scientific Ltd., UK) tightly placed on the skin of the subjects following instructions of the company. Each monitor had been turned on and placed in the laboratory for 10 mins prior to the sessions for signal stabilization. In addition, power output (**pw**) was also monitored every second using power meter pedals (Assioma, Favero Electronics SRL, Arcade TV, Italy).

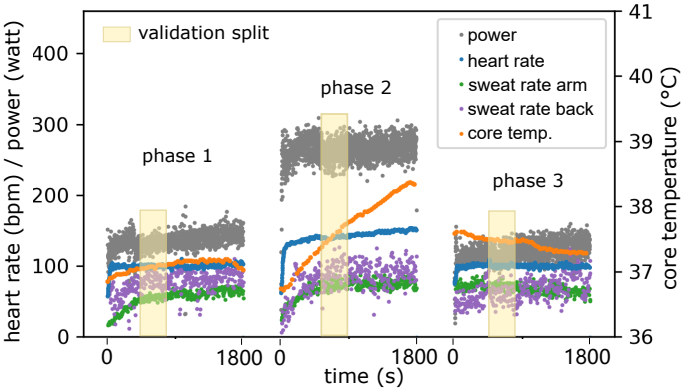

Fig. 1. Example sensor recordings in one exercise session and train/validation split.

III. Experimental Methods

A. Data Preprocessing

We use data collected across E1, E2, and E3 for further analysis, including biomarker recordings of $3 \times 1800s$ for each subject. Two major problems are addressed during signal preprocessing: (1) The measurements contain missing values due to data collection practices. (2) The sensors have different sampling rates (**hr**, **pw**: 1 Hz; **sra**, **srb**: 1/6 Hz; **ct**: 1/30 Hz).

For sweat rate recordings, values from some phases are entirely missing due to sensor failure. In that case, we fill the whole phase with the average sweat rate measurements of that phase from other subjects. For biomarkers that are partly missing in a phase, we carry out linear interpolation to fill in the missing values for that biomarker. Finally, we linearly interpolate **sra**, **srb**, and **ct** to 1 Hz so that all biomarker recordings have the same sampling rate.

B. Network Models

1) Support Vector Machine: We use a support vector machine (SVM) [14] classifier with radial basis function (RBF) kernel as the major baseline. ℓ_2 with coefficient 1 is used for regularization and the gamma coefficient in the RBF kernel is set to the inverse of the number of features.

2) Gated Recurrent Neural Network: We also use a RNN gated recurrent unit (GRU) [15] model with the formulations of the update of the reset and update gates (r_t, u_t) and the candidate and output states (c_t, h_t) as follows:

$$\begin{aligned} r_t &= \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r) \\ u_t &= \sigma(W_{xu}x_t + W_{hu}h_{t-1} + b_u) \\ c_t &= \tanh(W_{xc}x_t + r_t \odot (W_{hc}h_{t-1}) + b_c) \\ h_t &= (1 - u_t) \odot c_t + u_t \odot h_{t-1}, \end{aligned} \tag{1}$$

where $W_x, W_h \in \mathcal{R}^{M \times C}$ and $b \in \mathcal{R}^M$ are the learnable weight matrices and biases respectively, σ is the element-wise sigmoid function, $\odot$ is the hadamard product, M is the hidden layer size and C is the number of input channels.

We use 2-layer GRUs with 3 variants: GRU-$\{100, 200, 300\}$ representing hidden layer sizes of 100, 200, and 300 respectively. For training, we use the AdamW [16] optimizer with a

learning rate of 0.001 and a weight decay of 0.01. The GRU models are trained over 300 epochs.

3) XGBoost and Random Forest: We use extreme gradient boosting (XGBoost) [17] and random forest (RF) [18] for sensors ablation and importance study in Section IV-B and Section IV-C respectively. XGBoost is used as an additional baseline since it is known to be suitable for tabular data [19]. We choose an XGBoost model with 200 estimators and maximum depth of 8 for this study. We use a RF model to determine the permutation feature importance and impurity based feature importance so as to show the relative contribution of the biomarkers. The RF model ensembles 100 estimators and uses Gini impurity to measure the quality of a split. We set the max depth to 10 and min samples per split to 2.

C. Dataset Split Preparation

We split the preprocessed data from the 5 sensors into training and validation datasets, where the validation dataset contains 360 s of recordings from each phase. Each experiment is carried out 5 times with different random seeds that control the model initialization and the creation of the training and validation dataset. For each experiment, the validation data from all phases is collected from the recordings that start at the same random time point as shown by the yellow strips in Figure 1. This is to rule out any coincidental choice of a particularly easy to identify time period across subjects.

IV. Results

We first demonstrate the possibility of the person identification task by conducting dimension reduction and visualization with principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) method [20].

We first divide the data into non-overlapping time series. Each time series window consists of 30 seconds of data from all five sensor channels. We then flatten the time series data from each window and apply PCA to reduce the dimension to 20. Finally we apply the t-SNE method to further reduce the dimension to 2 for visualization on the plane.

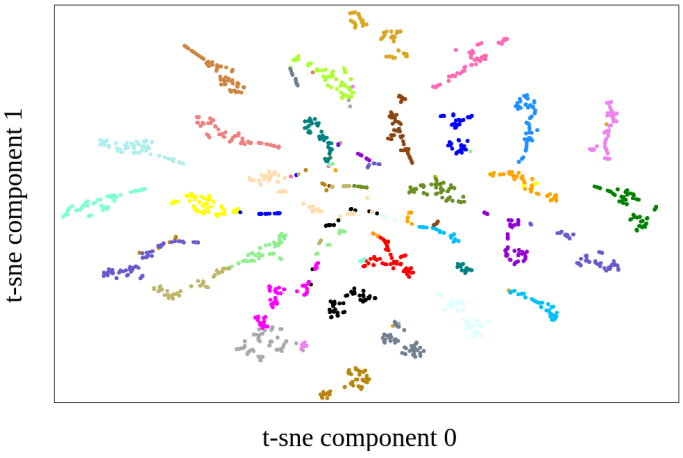

Fig. 2. T-SNE visualization on dimension reduced data. Only phase 1 data is used for ease of visualization. Each color represents a unique individual.

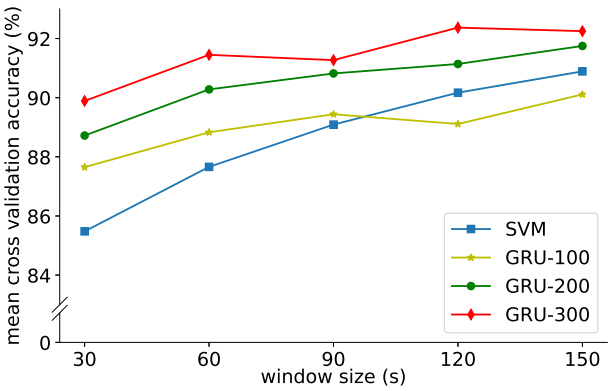

Fig. 3. Mean cross validation accuracies of person identification over 5 runs.

Figure 2 shows the dimension reduced visualization result. We see that the time series data from the same subject are distributed closely on the 2D plane, indicating that the biomarker measurements are sufficiently different across the subjects thereby allowing the possible distinction of a subject.

A. Comparison of Models on Person Identification

We then study whether it is possible to distinguish between the 30 subjects using ML methods with all sensor readings and the effect of using different window sizes (Ws) on the training and validation dataset. The biomarker recordings are split into 5-channel time series with Ws= $\{30s, 60s, 90s, 120s, 150s\}$ and stride length of 30 s. For SVM, each time series is further flattened to a 1D vector with the size of $5 \times$ Ws.

Figure 3 shows the 5-fold mean validation accuracies using different ML methods and Ws. The mean validation accuracy reaches $\sim$90% with the GRU-300 model and $\sim$85% with a simple SVM classifier both on 30 s of data. This suggests that less than a minute of biomarker collection is enough for the network to identify the exerciser with good accuracy. With Ws=150 s, all models show accuracy improvement in comparison with Ws=30 s. The GRU-300 model achieves the highest accuracy of $\sim$92%.

The validation accuracies from the models have a large standard deviation (see first row of Table II). We hypothesize that this is due to the large difference in accuracy results obtained from different validation time periods of the exercise session. Table I shows the mean and standard deviation of validation accuracies using SVM and GRU-200 with all sensor

TABLE I
Mean validation accuracies (%) $\pm$ standard deviation (%) of selected validation time period. Ws=30 s.

Validation time period (s)	***SVM accuracy (%)***	***GRU-200 accuracy (%)***
[0,360)	44.8	46.6 $\pm$ 0.5
[360,720)	86.1	89.8 $\pm$ 1.1
[720,1080)	**94.1**	**95.6 $\pm$ 1.0**
[1080,1440)	93.8	94.5 $\pm$ 0.3
[1440,1800)	84.9	85.5 $\pm$ 0.6

types on 5 validation time periods and Ws=30 s. The standard deviations are omitted for SVM since they are 0 when rounded to 2 decimal places. For every given validation time period, the standard deviations of mean validation accuracies for GRU-200 come solely from the model initialization.

We see that the validation data from the beginning period ($[0s, 360s)$) of all phases gives accuracies which are at least 39% lower than that of the other time periods. In general, both SVM and GRU-200 give higher accuracies for the middle period ($[360s, 1440s)$). The lower accuracy in the first validation time period of each exercise phase is probably due to transient changes in the biomarker measurements when exercise is started or resumed in later phases. This suggests that it is preferable to apply the models on the biomarkers measured a few minutes after the start of an exercise phase.

B. Ablation Study

We conduct ablation studies to investigate whether there are redundant sensors for the identification task. The parameters and training schedule of the SVM and GRU-200 are the same as in Section IV-A. Table II shows the mean and standard deviations of validation accuracies using SVM, XGBoost, and GRU-200 with different sensor combinations and Ws=30 s.

We find that incorporating more sensor readings is always advantageous for this person identification task, and all three models give the best accuracy with input from all five sensors. For the GRU-200 model, excluding **sra** gives the smallest accuracy drop (7.4%) in comparison to the accuracy from all sensors suggesting that this sensor is the least important for the model. For SVM and XGBoost, excluding **ct** gives the smallest accuracy drop (7.2% and 5.2% respectively). Excluding **srb** leads to the largest drop (11.6%) for the XGBoost method while excluding **hr** leads to the largest accuracy drop for both the SVM (17.5%) and GRU-200 (15.2%) models.

C. Sensor Importance

We use the RF model to study the relative contribution of the sensors. The baseline mean validation accuracy with all five sensors is 82.6%. The permutation feature importance for a certain feature is measured as the decrease of the RF classification accuracy from this baseline, when all the values of this feature are randomly shuffled. Another feature importance measure, the impurity based feature importance, reports the normalized decrease of impurity score brought by a specific feature. For both measures, the larger the value, the more important the feature.

TABLE II
ABLATION STUDIES ON SENSOR INPUTS USING WS=30 S.

Input sensors	Identification accuracy in % (mean ± std dev)		
	SVM	*XGBoost*	*GRU-200*
hr+pw+ct+sra+srb	**85.5 ± 8.7**	**82.8 ± 9.1**	**89.8 ± 6.3**
hr+pw+ct+sra	70.3 ± 11.1	71.2 ± 10.3	78.7 ± 9.9
hr+pw+ct+srb	76.5 ± 7.8	76.4 ± 8.8	82.4 ± 7.6
hr+ct+sra+srb	68.3 ± 11.9	72.9 ± 10.5	78.0 ± 12.0
hr+pw+sra+srb	78.3 ± 10.8	77.6 ±11.6	80.7 ± 10.2
pw+ct+sra+srb	68.0 ± 10.7	72.9 ± 8.9	74.6 ± 8.9
hr+pw+ct	51.9 ± 7.6	57.5 ± 7.8	63.7 ± 5.4
hr+ct	28.9 ± 2.8	28.4 ± 3.9	28.8 ± 4.8

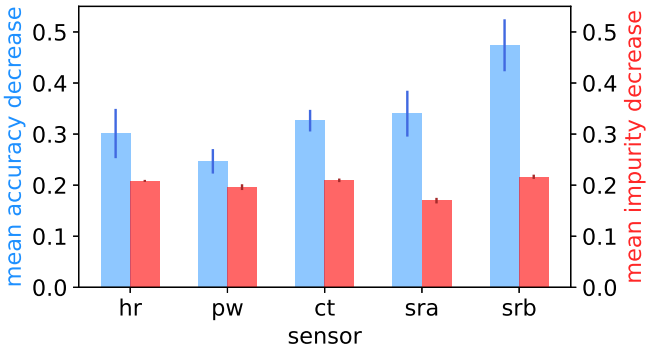

Fig. 4. Sensor importance scores. For both mean accuracy and Gini impurity measurements, a larger decrease means a larger importance of the sensor.

Figure 4 shows the sensor importance scores from the two measures for the five sensors. We see that all five sensors have non-negligible importance scores. **srb** is ranked the highest using both measures and matches the results of the XGBoost model in the ablation study. Together with results from the ablation study, we see that all five sensors are useful for this person identification task. Increasing the number of sensors from 4 to 5 leads to an increase in the accuracy from 5.2% to 17.5% with any of the ML methods used in this work.

V. Discussion and Conclusion

We present cycling exercise experiments on 30 subjects to determine if physiological biomarkers can be used for a person identification task. We test known machine learning methods including recurrent neural networks on collected physiological biomarkers including heart rate, biking power, core body temperature and sweat rate. The biomarker recordings are treated as time series inputs and despite the existence of sensory noise, the machine learning methods are able to robustly identify the person with over 92% accuracy with the use of all biomarkers by collecting data for 150 seconds time period. Even with measurements collected within a short time window of 30 s, the model achieved ∼90% accuracy. Ablation studies show that the accuracy decreases if any of the sensory input is absent. The networks can be deployed on mobile platforms for real-time processing of these time-varying signals [21], [22]. Results show that we can achieve 22 inferences/sec when running the GRU-200 model on the Jetson Nano using the low power mode of 5 W. This shows that we can easily integrate this module in a wearable device. The person identification output could potentially be used for building user profiles and for downstream tasks such as exercise planning and recommendation. We also carried out preliminary experiments to see if eliminating the phases where entire recordings of 1 or 2 sensors are missing due to recording caveats, will affect the task accuracy. Our preliminary results from SVM ($85.0 \pm 9.0\%$ for Ws=30 s and $91.1 \pm 5.2\%$ for Ws=150 s) showed that the accuracies were similar.

REFERENCES

[1] K. Bayoumy, M. Gaber, A. Elshafeey, O. Mhaimeed, E. H. Dineen, F. A. Marvel, S. S. Martin, E. D. Muse, M. P. Turakhia, K. G. Tarakji, and M. B. Elshazly, "Smart wearable devices in cardiovascular care: where we are and how to move forward," *Nature Reviews Cardiology*, vol. 18, no. 8, pp. 581–599, aug 2021.

[2] E. Guillodo, C. Lemey, M. Simonnet, M. Walter, E. Baca-García, V. Masetti, S. Moga, M. Larsen, H. Network, J. Ropars *et al.*, "Clinical applications of mobile health wearable–based sleep monitoring: systematic review," *JMIR mHealth and uHealth*, vol. 8, no. 4, p. e10733, 2020.

[3] D. R. Seshadri, R. T. Li, J. E. Voos, J. R. Rowbottom, C. M. Alfes, C. A. Zorman, and C. K. Drummond, "Wearable sensors for monitoring the physiological and biochemical profile of the athlete," *npj digital medicine*, vol. 2, no. 1, pp. 1–16, 2019. [Online]. Available: https://doi.org/10.1038/s41746-019-0150-9

[4] F. Sabry, T. Eltaras, W. Labda, F. Hamza, K. Alzoubi, and Q. Malluhi, "Towards on-device dehydration monitoring using machine learning from wearable device's data," *Sensors*, vol. 22, no. 5, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/5/1887

[5] J. Choi, R. Ghaffari, L. B. Baker, and J. A. Rogers, "Skin-interfaced systems for sweat collection and analytics," *Science advances*, vol. 4, no. 2, p. eaar3921, 2018.

[6] A. J. Bandodkar, W. J. Jeang, R. Ghaffari, and J. A. Rogers, "Wearable sensors for biochemical sweat analysis," *Annual Review of Analytical Chemistry*, vol. 12, no. 1, pp. 1–22, 2019, pMID: 30786214. [Online]. Available: https://doi.org/10.1146/annurev-anchem-061318-114910

[7] K. A. Barnes, M. L. Anderson, J. R. Stofan, K. J. Dalrymple, A. J. Reimel, T. J. Roberts, R. K. Randell, C. T. Ungaro, and L. B. Baker, "Normative data for sweating rate, sweat sodium concentration, and sweat sodium loss in athletes: An update and analysis by sport," *Journal of Sports Sciences*, vol. 37, no. 20, pp. 2356–2366, 2019.

[8] S. Wang, C. Lafaye, M. Saubade, C. Besson, J. M. Margarit-Taulé, V. Gremeaux, and S.-C. Liu, "Predicting hydration status using machine learning models from physiological and sweat biomarkers during endurance exercise: a single case study," *IEEE Journal of Biomedical and Health Informatics*, 2022.

[9] P. Thiam, H. Hihn, D. A. Braun, H. A. Kestler, and F. Schwenker, "Multi-modal pain intensity assessment based on physiological signals: A deep learning perspective," *Frontiers in Physiology*, vol. 12, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fphys.2021.720464

[10] T. Beer, B. Eini-Porat, S. Goodfellow, D. Eytan, and U. Shalit, "Using deep networks for scientific discovery in physiological signals," in *Proceedings of the 5th Machine Learning for Healthcare Conference*, ser. Proceedings of Machine Learning Research, F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, Eds., vol. 126. PMLR, 07–08 Aug 2020, pp. 685–709. [Online]. Available: https://proceedings.mlr.press/v126/beer20a.html

[11] I. Jayarathne, M. Cohen, and S. Amarakeerthi, "Person identification from EEG using various machine learning techniques with inter-hemispheric amplitude ratio," *PLOS ONE*, vol. 15, no. 9, pp. 1–24, 09 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0238872

[12] D. Jyotishi and S. Dandapat, "An lstm-based model for person identification using ecg signal," *IEEE Sensors Letters*, vol. 4, no. 8, pp. 1–4, 2020.

[13] J. Hopker, D. Coleman, and L. Passfield, "Changes in cycling efficiency during a competitive season," *Medicine & Science in Sports & Exercise*, vol. 41, no. 4, pp. 912–919, 2009.

[14] C. Cortes and V. N. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 2004.

[15] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 103–111. [Online]. Available: https://aclanthology.org/W14-4012

[16] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[17] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 785–794. [Online]. Available: https://doi.org/10.1145/2939672.2939785

[18] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 10 2001.

[19] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," in *8th ICML Workshop on Automated Machine Learning (AutoML)*, 2021. [Online]. Available: https://openreview.net/forum?id=vdgtepS1pV

[20] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[21] S. Wang, Y. Hu, J. Burgués, S. Marco, and S.-C. Liu, "Prediction of gas concentration using gated recurrent neural networks," in *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, 2020, pp. 178–182.

[22] N. LeBow, B. Rueckauer, P. Sun, M. Rovira, C. Jiménez-Jorquera, S.-C. Liu, and J. M. Margarit-Taulé, "Real-time edge neuromorphic tasting from chemical microsensor arrays," *Frontiers in Neuroscience*, vol. 15, 2021.