Neural Computing and Applications (2022) 34:18933–18951
https://doi.org/10.1007/s00521-022-07505-w

ORIGINAL ARTICLE

An abstract parabolic system-based physics-informed long short-term memory network for estimating breath alcohol concentration from transdermal alcohol biosensor data

Clemens Oszkinat[1] · Susan E. Luczak[2] · I. Gary Rosen[1]

Received: 5 April 2022 / Accepted: 1 June 2022 / Published online: 26 June 2022

Abstract

The problem of estimating breath alcohol concentration based on transdermal alcohol biosensor data is considered. Transdermal alcohol concentration provides a promising alternative to classical methods such as breathalyzers or drinking diaries. A physics-informed long short-term memory (LSTM) network with covariates for the solution of the estimation problem is developed. The data-driven nature of an LSTM is augmented with a first-principles physics-based population model for the diffusion of ethanol through the epidermal layer of the skin. The population model in an abstract parabolic framework appears as part of a regularization term in the loss function of the LSTM. While learning, the model is encouraged to both fit the data and to produce physically meaningful outputs. To deal with the high variation observed in the data, a mechanism for the uncertainty quantification of the estimates based on a recently discovered relation between Monte-Carlo dropout and Bayesian learning is used. The physics-based population model and the LSTM are trained and tested using controlled laboratory collected breath and transdermal alcohol data collected in four sessions from 40 orally dosed participants (50% female, ages 21–33 years, 35% BMI above 25.0) resulting in 256 usable drinking episodes partitioned into training and testing sets. Body measurement (e.g., BMI, hip to waist ratio, etc.), personal (e.g., sex, age, race, etc.), drinking behavior (e.g., frequent, rarely, etc.), and environmental (e.g., temperature, humidity, etc.) covariates were also collected from participants. The importance of various covariates in the estimation is investigated using Shapley values. It is shown that the physics-informed LSTM network can be successfully applied to drinking episodes from both the training and test sets and that the physics-based information leads to better generalization ability on new drinking episodes with the uncertainty quantification yielding credible bands that effectively capture the true signal. Compared to two machine learning models from previous studies, the proposed model reduces relative L_2 error in estimated breath alcohol concentration by 58 and 72% and relative peak error by 33 and 76%.

Keywords Physics-informed machine learning · LSTM · Transdermal alcohol concentration · Biosensor data

1 Introduction

Alcohol researchers and clinicians alike have long been interested in qualitatively and quantitatively measuring and tracking alcohol consumption. In particular, near-continuous measurements are considered important as they can be used for the monitoring of blood alcohol, the discovery of addiction problems and the examination of naturalistic drinking patterns.

However, producing meaningful quantitative measures of alcohol consumption in naturalistic settings is a challenging task. Typically, this is based on the use of a breath alcohol analyzer or a drinker's self-report. Unfortunately,

✉ I. Gary Rosen
grosen@usc.edu

Clemens Oszkinat
oszkinat@usc.edu

Susan E. Luczak
luczak@usc.edu

[1] Department of Mathematics, University of Southern California, Los Angeles, CA 90089, USA

[2] Department of Psychology, University of Southern California, Los Angeles, CA 90089, USA

both methods have their shortcomings: Obtaining deep lung samples (alveolar air) needed for accurate results can be difficult and often, alcohol contained in the mouth after drinking contaminates the results [1]. Also, the procedure does obviously not allow for continuous measurements; the intense exhaling cannot be repeated frequently. Clearly, this method is not well suited for studying drinking in a naturalistic setting. Self-reports, on the other hand, are often inaccurate as well. It is difficult for people to judge what quantities of alcohol they have consumed in a particular situation. If an alcoholic drink was prepared by someone else, e.g., a drink at a bar, this process is impractical, and even worse, it is known that alcohol directly affects the memory function of the brain [2]; hence, a drinker's self-report will likely lack the required reliability.

After the consumption of alcohol, the human liver metabolizes most of the alcohol. Some of the alcohol is exhaled through breath, some leaves the body through urine, and about 1% of the alcohol diffuses through the skin [3]. The excretion of ethanol through perspiration has been known since the 1930s [4–6]. Hence, measuring transdermal alcohol creates a possible alternative for the tracking of alcohol consumed. Recently, biosensors have been developed that are able to measure transdermal alcohol level or concentration (TAC) [7–9]. These sensors primarily function as a fuel cell: In a redox reaction, 4 electrons are produced for every ethanol molecule. This generates a current that is measured and converted into TAC.

The availability of TAC measuring devices allows for near-continuous measurements of alcohol consumption and helps researchers to gain insight into alcohol metabolism and drinking behavior in a naturalistic setting. In addition, TAC sensors collect the data passively, i.e., contrary to self-reports or breath alcohol analyzers no active participation of the subject is required. In this context, those devices can be used for forensic applications such as DUI (driving under the influence: the offense of driving a vehicle while impaired by alcohol or drugs) monitoring and the underlying mechanism can potentially also be extended to other drugs. Recent interest in wearable computers could lead to TAC biosensors being integrated into these devices to reach the consumer market.

Researchers interested in drinking behavior and alcohol consumption typically base their studies on breath alcohol concentration (BrAC) or blood alcohol concentration (BAC), which have been shown to reasonably agree [10]. Hence, in order to make TAC sensors useful for alcohol research, the need to convert TAC signals to BAC/BrAC signals arises. Unfortunately, the direct conversion of TAC to BAC/BrAC has proven to be challenging due to a number of confounding factors. Differences between devices, a subject's drinking behavior, and varying environmental conditions such as temperature and humidity lead to variations in the measurements. Intra- and inter-individual variations are additional sources that pose further challenges to the direct conversion from TAC to BAC/BrAC. The porosity, thickness, tortuosity, hydration, and vasodilation of an individual's skin are important factors in the functional relationship between BAC/BrAC and TAC.

In the last decade, our research team and other groups have worked on different approaches to overcome the difficulties in TAC-based BrAC estimation. Some early efforts used deterministic models [11–17] for the relationship between BAC/BrAC and TAC. Some of them were based on regression models [12], while others modeled the transport of alcohol from the blood through the epidermis by a one-dimensional diffusion equation with unknown parameters. This diffusion model takes the form of a partial differential equation (PDE), a class of equations that are ubiquitous in biological and engineering models. The idea to approach the BrAC estimation using machine learning techniques is relatively new. In [18], the authors used a tree-based ensemble regression algorithm to create estimates of BrAC from transdermal data. A treatment by our research group employed a covariate-dependent Hidden Markov Model to account for body-level covariates such as height and age which was shown to yield meaningful BrAC estimates together with error bands to quantify the associated uncertainty [19]. We note that this was one of the first attempts to incorporate covariate information into the estimation of BrAC from TAC. In another study, [20], we used a hybrid approach combining the aforementioned PDE model with the data-driven nature of machine learning using a physics-informed generative adversarial network (GAN). Such a hybrid model is especially interesting in the context of the alcohol biosensor problem. Although there is a first-principles physics model, it is known that such a model is not able to fully explain the observed variation in the relation between TAC and BrAC. On the other hand, available data sets with TAC/BrAC pairs only contain about a hundred data pairs due to the difficult and time-consuming nature of gathering human subject data, compare [21–23]. This scarcity of data poses a problem for the training of complex machine learning models. Hence, in the present work, we want to maintain the idea of a hybrid model to enrich the data-driven qualities of machine learning with the a priori physics information. But other than in [20], where the input/output PDE model was considered on the full domain and where a GAN was used, here we want to make use of recent advances in time series machine learning in combination with an abstract evolutionary, semigroup theory-based formulation of the PDE model. To this end, we propose a physics-informed long short-term memory (LSTM)

network. In this model, the data-driven nature of an LSTM is augmented with a first-principles physics-based model for the diffusion of ethanol through the epidermal layer of the skin. This physical model takes the role of a regularization term in the loss function of the LSTM. That way, while learning, the model is encouraged to both fit the data and to produce physically meaningful outputs. Moreover, due to the high level of variation observed in the data, we also propose a mechanism for the uncertainty quantification of the estimates that is based on a recently discovered relationship between Monte-Carlo dropout and Bayesian learning.

The remainder of the paper is organized as follows: In Sect. 2, we introduce the physical model for the transport of alcohol through the epidermal layer of the skin. Also in this section, we reformulate the model as an abstract parabolic evolution equation to make it readily amenable to the training process. Then, in Sect. 3, we describe the LSTM network that is used in this work and we show how the physics information is included in the network training process to obtain a physics-informed LSTM model. In Sect. 4, we briefly explain the network architecture that we use to handle both time-dependent and time-independent covariates, followed in Sect. 5 by an outline of the approach used to capture the uncertainty in our BrAC estimates. Finally, in Sect. 6, we present some examples of our numerical studies. We conclude with a discussion of our results and some final remarks.

2 A diffusion model for transdermal alcohol transport

The process of ethanol diffusion through the epidermal layer of the skin has been carefully studied over the past years, and several, mostly similar, physical models have been developed to capture the physiological processes. Here, we use a distributed parameter diffusion model that was described in [24] and takes the form

$$\frac{\partial x}{\partial t}(t,\eta) = q_1 \frac{\partial^2 x}{\partial \eta^2}(t,\eta), \quad t > 0,\ \eta \in (0,1), \tag{1}$$

$$\frac{dw}{dt}(t) = q_3 \frac{\partial x}{\partial \eta}(t,0) - q_4 w(t), \quad t > 0, \tag{2}$$

$$x(t,0) = w(t), \quad t > 0, \tag{3}$$

$$q_1 \frac{\partial x}{\partial \eta}(t,1) = q_2 u(t), \quad t > 0, \tag{4}$$

$$w(0) = w_0, \quad x(0,\eta) = \phi_0(\eta), \quad \eta \in (0,1) \tag{5}$$

$$y(t) = w(t), \quad t \geq 0. \tag{6}$$

In this model, $x(t,\eta)$ is the concentration of ethanol in the epidermal layer at time t and depth $\eta \in [0,1]$ and $w(t)$ is the concentration of ethanol in the sensor's collection chamber at time t. The TAC output is denoted by $y(t)$. In the case of the largely dimensionless system considered above, $w(t)$ and $y(t)$ coincide. The BrAC input to the model is denoted by $u(t)$. The dimensionless parameter q_1 models the diffusivity of the skin and the dimensionless q_2 describes the alcohol gain from the dermal layer of the skin. Both parameters are unknown and un-measurable; they are subject-dependent physiological parameters. The parameters q_3 and q_4 are device-dependent. In what follows, we will assume the parameters $\boldsymbol{q} = [q_1, q_2, q_3, q_4]$ to be random and we estimate their distributions in order to use them in the above model.

The model describes a (linear) relationship between the BrAC u as the input of the model and the TAC y as the output of the model. However, since the model is given in the time-space domain, it is not directly accessible to a time series machine learning model. Moreover, we need to discretize the model appropriately, and we need to account for the randomness of the parameters $\boldsymbol{q}$. Hence, in the following, we rewrite the initial-boundary problem (1)–(6) as an abstract parabolic system in a Gelfand triple of Hilbert (and later Bochner) spaces. Then, we show how the random parameters $\boldsymbol{q}$ can essentially be considered as additional spatial variables, and how the resulting model can be discretized. Here, we follow the steps that can be found for example in [24] and [25].

First, we consider a finite time horizon, $T > 0$, and we assume a zero-order hold input, that is $u(t) = u_k, t \in [k\tau, (k+1)\tau), k = 0, 1, 2, \ldots$ with τ denoting a sampling time. Such an assumption is not unrealistic in this application, as biosensors have a sampling time, and breath alcohol concentration changes slowly over the course of hours. Analogously, we define $x_k = x_k(\eta) = x(k\tau, \eta), w_k = w(k\tau)$, and $y_k = y(k\tau), k = 0, 1, \ldots, K$, with $K\tau = T$. Then, we consider the system (1)–(6) on the intervals $[k\tau, (k+1)\tau]$ for $k = 0, 1, 2, \ldots, K$. With the change of variables $v(t,\eta) = x(t,\eta) - \xi(\eta)u_k$ where $\xi(\eta) = \frac{q_2}{q_1}\eta$, we obtain the system

$$\frac{\partial v}{\partial t}(t,\eta) = q_1 \frac{\partial^2 v}{\partial \eta^2}(t,\eta), \quad k\tau < t < (k+1)\tau,\ \eta \in (0,1), \tag{7}$$

$$\frac{dw}{dt}(t) = q_3 \frac{\partial v}{\partial \eta}(t,0) - q_4 w(t) + \frac{q_3 q_2}{q_1} u_k, \quad k\tau < t < (k+1)\tau, \tag{8}$$

$$v(t,0) = w(t), \quad k\tau < t(k+1)\tau, \tag{9}$$

$$q_1 \frac{\partial v}{\partial \eta}(t,1) = 0, \quad k\tau < t < (k+1)\tau, \tag{10}$$

$$v(k\tau,\cdot) = x(k\tau,\cdot) - \xi(\cdot)u_k = x_k - \xi u_k. \tag{11}$$

Let Q be a compact subset of the positive orthant of $\mathbb{R}^4$. For $\boldsymbol{q} = [q_1, q_2, q_3, q_4] \in Q$ we define $H_q = \mathbb{R} \times L^2(0,1)$ with the inner product $\langle(\theta_1,\phi_1),(\theta_2,\phi_2)\rangle_q = \frac{q_1}{q_3}\theta_1\theta_2 + \int_0^1 \phi_1(\eta)\phi_2(\eta)d\eta$. Define further the Hilbert space $V = \{(\theta,\phi) \in H_q : \phi \in H^1(0,1), \theta = \phi(0)\}$ with the inner product $\langle(\phi_1(0),\phi_1),(\phi_2(0),\phi_2)\rangle_V = \phi_1(0)\phi_2(0) + \langle\phi_1',\phi_2'\rangle_{L^2(0,1)}$. Here, $\langle\cdot,\cdot\rangle_{L^2(0,1)}$ denotes the standard $L_2(0,1)$ inner product.

Using standard arguments, we obtain the dense and continuous embeddings $V \hookrightarrow H_q \hookrightarrow V^*$ and that these embeddings are uniformly bounded with respect to $\boldsymbol{q} \in Q$. Further, for each $\boldsymbol{q} \in Q$, we define the bilinear form $a(\boldsymbol{q};\cdot,\cdot) : V \times V \to \mathbb{R}$ by

$$a(\boldsymbol{q};\hat{\phi}_1,\hat{\phi}_2) = \frac{q_1 q_4}{q_3}\phi_1(0)\phi_2(0) + q_1 \int_0^1 \phi_1'(\eta)\phi_2'(\eta)d\eta \tag{12}$$

for $\phi_1, \phi_2 \in V$ and where $\hat{\phi}_1 = (\phi_1(0), \phi_1)$, $\hat{\phi}_2 = (\phi_2(0), \phi_2)$. Since Q was assumed to be compact, the bilinear form $a(\boldsymbol{q};\cdot,\cdot)$ is uniformly bounded, coercive, and continuously dependent on $\boldsymbol{q} \in Q$. See [24] for more details.

Now, we define the operator $A(\boldsymbol{q}) : \mathrm{Dom}(A(\boldsymbol{q})) \subset H_q \to H_q$ by $\langle A(\boldsymbol{q})\hat{\phi},\hat{\psi}\rangle_{V^*,V} = -a(\boldsymbol{q};\hat{\phi},\hat{\psi})$ where $\mathrm{Dom}(A(\boldsymbol{q})) = \{\hat{\phi} = (\phi(0),\phi) \in V : \phi \in H^2(0,1)\}$. Note that $\mathrm{Dom}(A(\boldsymbol{q}))$ does not depend on $\boldsymbol{q}$. With that, one can show that

$$A(\boldsymbol{q})\hat{\phi} = A(\boldsymbol{q})(\phi(0),\phi) = (q_3\phi'(0) - q_4\phi(0), q_1\phi'') \tag{13}$$

for $\hat{\phi} \in \mathrm{Dom}(A(\boldsymbol{q}))$. Further, $A(\boldsymbol{q})$ is densely defined on H_q, regularly dissipative and self-adjoint and thus, $A(\boldsymbol{q})$ is an infinitesimal generator of a uniformly exponentially stable, self-adjoint, analytic semigroup of bounded linear operators, $\{e^{A(\boldsymbol{q})t} : t \geq 0\}$, on H_q. For details, we refer to [26, 27].

With this, the system (7)–(11) takes the form $\frac{d\hat{v}}{dt}(t) = A(\boldsymbol{q})\hat{v}(t) + (\frac{q_3 q_2}{q_1}, 0)u_k$ with $\hat{v}(k\tau) = (w_k, x_k - \xi u_k)$ in the interval $k\tau < t < (k+1)\tau$. Define $\hat{x_k} = (w_k, x_k)$ and the operator $\hat{A}(\boldsymbol{q}) = e^{A(\boldsymbol{q})\tau} \in \mathscr{L}(H_q, H_q)$. Further, define the operator $B(\boldsymbol{q}) = \left(\frac{q_3 q_2}{q_1}, 0\right) - A(\boldsymbol{q})(0,\xi)$. Then set

$$\begin{aligned}\hat{B}(\boldsymbol{q}) &= \int_0^\tau e^{A(\boldsymbol{q})s}B(\boldsymbol{q})ds = A(\boldsymbol{q})^{-1}e^{A(\boldsymbol{q})s}B(\boldsymbol{q})|_0^\tau \\ &= (\hat{A}(\boldsymbol{q}) - I)A(\boldsymbol{q})^{-1}B(\boldsymbol{q}) \\ &= (I - \hat{A}(\boldsymbol{q}))\left((0,\xi) - A(\boldsymbol{q})^{-1}\left(\frac{q_3 q_2}{q_1}, 0\right)\right)\end{aligned} \tag{14}$$

$\in \mathscr{L}(\mathbb{R}, H_q)$. With these definitions, we obtain the discrete-time state space form of the model as

$$\begin{aligned}\hat{x}_{k+1} &= (w_{k+1}, x_{k+1}) = (w((k+1)\tau), x((k+1)\tau,\cdot)) \\ &= \hat{v}((k+1)\tau) + (0,\xi)u_k \\ &= e^{A(\boldsymbol{q})\tau}(w_k, x_k - \xi u_k) \\ &\quad + \int_0^\tau e^{A(\boldsymbol{q})s}\left(\frac{q_3 q_2}{q_1}, 0\right)ds\, u_k + (0,\xi)u_k \\ &= \hat{A}(\boldsymbol{q})\hat{x}_k + (I - \hat{A}(\boldsymbol{q}))(0,\xi)u_k \\ &\quad + A(\boldsymbol{q})^{-1}(\hat{A}(\boldsymbol{q}) - I)\left(\frac{q_3 q_2}{q_1}, 0\right)u_k \\ &= \hat{A}(\boldsymbol{q})\hat{x}_k + \hat{B}(\boldsymbol{q})u_k.\end{aligned} \tag{15}$$

Defining further the output operator $\hat{C} \in \mathscr{L}(H_q, \mathbb{R})$ as $\hat{C}(\theta,\phi) = \theta$ for $(\theta,\phi) \in H_q$, our discrete time input/output model becomes

$$\hat{x}_{k+1} = \hat{A}(\boldsymbol{q})\hat{x}_k + \hat{B}(\boldsymbol{q})u_k, \quad k = 0,1,2,\ldots \tag{16}$$

$$\hat{x}_0 = (w_0, \phi_0), \tag{17}$$

$$y_k = \hat{C}\hat{x}_k, \quad k = 0,1,2,\ldots. \tag{18}$$

Assuming zero initial conditions for the epidermal layer of the skin and the collection chamber of the sensor, that is $\hat{x}_0 = (w_0,\phi_0) = (0,0)$, the output y of the system takes the form of a discrete time convolution of the input, u, with a filter, $h(\boldsymbol{q})$, as

$$\begin{aligned}y_k &= \sum_{j=0}^{k-1} \hat{C}\hat{A}(\boldsymbol{q})^{k-j-1}\hat{B}(\boldsymbol{q})u_j \\ &= \sum_{j=0}^{k-1} h_{k-j-1}(\boldsymbol{q})u_j, \quad k = 0,1,2,\ldots.\end{aligned} \tag{19}$$

Here, the filter is $h_i(\boldsymbol{q}) = \hat{C}\hat{A}(\boldsymbol{q})^i\hat{B}(\boldsymbol{q}), i = 0,1,2,\ldots$. Using the Trotter-Kato approximation theorem from linear semigroup theory [28] it can be shown that the mapping $\boldsymbol{q} \mapsto h_i(\boldsymbol{q})$ is continuous, uniformly in $\boldsymbol{q}$ and i for $\boldsymbol{q} \in Q$ and $i = 0,1,2,\ldots,K$.

Now, shifting from the assumption that $\boldsymbol{q}$ is deterministic to the idea of $\boldsymbol{q}$ being a random vector, we denote this random vector as φ and assume that φ has support in the cube $\prod_{i=1}^4 [a_i, b_i]$ for some $-\infty < \bar{a} < a_i < b_i < \bar{b} < \infty$.

Define the vectors $\mathbf{a} = [a_i]_{i=1}^4$, $\mathbf{b} = [b_i]_{i=1}^4$ and let $\Theta \subset \mathbb{R}^4$ be a compact parameter set. Assume further that q is distributed according to an absolutely continuous cumulative density function, $F(q; [\mathbf{a}, \mathbf{b}, \rho])$ or equivalently by a push forward measure $\pi = \pi([\mathbf{a}, \mathbf{b}, \rho])$ for $\rho \in \Theta$. We set $\theta = (\mathbf{a}, \mathbf{b}, \rho)$ and write $\pi(\theta) \sim f(\theta)$, where $f(\theta) = f(\cdot; \theta)$ is a density function. Based on this, we define the Bochner spaces $\mathscr{V} = L^2_\pi(Q; V)$ and $\mathscr{H} = L^2_\pi(Q; H_q)$ and obtain the Gelfand triple of spaces $\mathscr{V} \hookrightarrow \mathscr{H} \hookrightarrow \mathscr{V}^*$ where $\mathscr{V}^*$ denotes the dual of $\mathscr{V}$. Further, we define the bilinear form $a(\cdot, \cdot)$ on $\mathscr{V} \times \mathscr{V}$ as well as the operator $\mathscr{A} \in \mathscr{L}(\mathscr{V}, \mathscr{V}^*)$ by

$$\begin{aligned} a(\hat{\phi}, \hat{\psi}) &= \mathbb{E}_\pi\Big(a\Big(\boldsymbol{q}; \hat{\phi}(\boldsymbol{q}), \hat{\psi}(\boldsymbol{q})\Big)\Big) \\ &= \int_Q a\Big(\boldsymbol{q}; \hat{\phi}(\boldsymbol{q}), \hat{\psi}(\boldsymbol{q})\Big) \mathrm{d}\pi(\boldsymbol{q}) = -\langle \mathscr{A}\hat{\phi}, \hat{\psi} \rangle_{\mathscr{V}^*, \mathscr{V}} \end{aligned} \tag{20}$$

for $\hat{\phi}, \hat{\psi} \in \mathscr{V}$. As in the deterministic case, the operator $\mathscr{A}$ is regularly dissipative and self-adjoint and thus, $\mathscr{A}$ is an infinitesimal generator of a uniformly exponentially stable, analytic semigroup of bounded, self-adjoint, linear operators, $\{\mathscr{T}(t) : t \geq 0\}$, on $\mathscr{H}$ (see [26, 27, 29]). Analogously, we define the operator $\mathscr{B} \in \mathscr{L}(\mathbb{R}, \mathscr{H})$ by $\langle \mathscr{B}u, \hat{\psi} \rangle = \mathbb{E}_\pi\Big(\langle B(\boldsymbol{q})u, \hat{\psi} \rangle_{\boldsymbol{q}}\Big)$, $u \in \mathbb{R}$ and set $\hat{\mathscr{A}} = \mathscr{T}(\tau) \in \mathscr{L}(\mathscr{H}, \mathscr{H})$, $\hat{\mathscr{B}} = \mathscr{A}^{-1}(\hat{\mathscr{A}} - \mathscr{I})\mathscr{B} \in \mathscr{L}(\mathbb{R}, \mathscr{H})$, where $\mathscr{I}$ is the identity on $\mathscr{H}$. Lastly, we define the output operator $\hat{\mathscr{C}}$ by $\hat{\mathscr{C}}\hat{\phi} = \mathbb{E}_\pi(\hat{C}\hat{\phi}) \in \mathscr{L}(\mathscr{H}, \mathbb{R})$. With that, we obtain the discrete-time system

$$\hat{x}_{k+1} = \hat{\mathscr{A}}\hat{x}_k + \hat{\mathscr{B}}u_k \tag{21}$$

$$y_k = \hat{\mathscr{C}}x_k. \tag{22}$$

Once again, the output y_k can then be written as a convolution as

$$y_k = \sum_{j=0}^{k-1} \hat{\mathscr{C}}\hat{\mathscr{A}}^{k-j-1}\hat{\mathscr{B}}u_j = \sum_{j=0}^{k-1} h_{k-j-1}u_j, \quad k = 0, 1, 2, \ldots. \tag{23}$$

At this point, the model is in state space form, time-discrete, and the randomness of q has been taken into account. However, the filter h in (23) involves the semigroup $\{\mathscr{T}(t) : t \geq 0\}$ that is defined on the infinite dimensional Hilbert space h. Thus, as a last step, we need to apply finite dimensional approximation. To this end, we construct a series of approximating finite-dimensional subspaces $\mathscr{V}^N$ of $\mathscr{V}$ as well as sequences of operators $\hat{\mathscr{A}}^N, \hat{\mathscr{B}}^N$, and $\hat{\mathscr{C}}^N$ that approximate $\hat{\mathscr{A}}, \hat{\mathscr{B}}$, and $\hat{\mathscr{C}}$, respectively. Here, by N we denote the multi-index $N = (n, m_1, m_2, m_3, m_4)$. As before, we assume that the random parameters q_i are supported in $[a_i, b_i]$ for $i = 1, \ldots, 4$. Partition $[a_i, b_i]$ into m_i subintervals of equal width for $i = 1, \ldots, 4$ and denote the characteristic function on the j-th subinterval by $\chi_j^{m_i}$ for $j = 1, \ldots, m_i$. For $n = 1, 2, \ldots$ let $\{\phi_j^n\}_{j=0}^n$ be the standard linear B-splines on [0, 1] with respect to the uniform mesh $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$ and set $\hat{\phi}_j^n = (\phi_j^n(0), \phi_j^n) \in V$. With J we denote a multi-index $J = (j_0, j_1, \ldots, j_4)$ with $j_0 \in \{0, 1, \ldots, n\}$ and $j_i \in \{1, 2, \ldots, m_i\}, i = 1, 2, 3, 4$. We then set $\Phi_J^N = \hat{\phi}_{j_0}^n \prod_{i=1}^4 \chi_{j_i}^{m_i}$ and let $\mathscr{V}^N = \mathrm{span}_J\{\Phi_J^N\}$. Then, we let $\mathscr{P}^N : h \to \mathscr{V}^N$ be the orthogonal projection of h onto $\mathscr{V}^N$. Using standard arguments from spline theory and piecewise constant approximation in L^2, one can see that $\mathscr{P}^N$ converges strongly to the identity in both h and $\mathscr{V}$. This shows that the sequence $\mathscr{V}^N$ approximates $\mathscr{V}$ in an appropriate way. For the operator $\mathscr{A}$ we use a standard Galerkin approach to come up with an approximation $\mathscr{A}^N$, and set $\hat{\mathscr{A}}^N = \exp(\mathscr{A}^N\tau) \in \mathscr{L}(\mathscr{V}^N, \mathscr{V}^N)$, for details see [30]. Based on this, we set $\hat{\mathscr{B}}^N = (\mathscr{A}^N)^{-1}\big(\hat{\mathscr{A}}^N - \mathscr{I}^N\big)\mathscr{P}^N\mathscr{B} \in \mathscr{L}(\mathbb{R}, \mathscr{V}^N)$ and $\hat{\mathscr{C}}^N = \hat{\mathscr{C}}\mathscr{P}^N$. Standard arguments from linear semigroup theory can be used to show the convergence of the approximating operators. Finally, we can write

$$\begin{aligned} y_k^N &= \sum_{j=0}^{k-1} \hat{\mathscr{C}}^N(\hat{\mathscr{A}}^N)^{k-j-1}\hat{\mathscr{B}}^N u_j \\ &= \sum_{j=0}^{k-1} h_{k-j-1}^N u_j, \quad k = 0, 1, 2, \ldots \end{aligned} \tag{24}$$

for the output y^N that only depends on the input u. This is precisely the form we are going to use for the physics information in the neural network (see the next section).

3 A physics-informed long short-term memory network

We develop a BrAC estimator based on a Long Short-Term Memory Network as developed by Hochreiter and Schmidhuber in [31]. These networks are a class of recurrent neural networks and are especially useful for time-series data as they are able to process sequences of data and remember temporal features over long time spans.

The whole network can be thought of as T LSTM units stacked next to each other, where T denotes the number of time steps. In a single unit, there is a cell state c as well as a forget gate f, an input gate i, and an output gate o that are used to regulate the flow of information through the network. Denoting the time by the subscript t, the equations for these gates are given by

$$f_t = \sigma(W_f y_t + U_f h_{t-1} + b_f) \tag{25}$$

$$i_t = \sigma(W_i y_t + U_i h_{t-1} + b_i) \tag{26}$$

$$o_t = \sigma(W_o y_t + U_o h_{t-1} + b_o). \tag{27}$$

Here, h_t denotes the so-called hidden state that is passed from one time step $t-1$ to the next, t, and y_t denotes the input to the network at time t, in our case $y = (y_1, \ldots, y_T)$ is a TAC signal or a vector of multiple time-dependent covariates, for example TAC and heart rate. The quantity f_t denotes the forget gate. Here, the current input y_t is weighted by a matrix W_f and the previous hidden state h_{t-1} is weighted by a matrix U_f. Then, a bias term b_f is added and a sigmoid function $\sigma(x) = 1/(1+e^{-x})$ is applied to the result. Similar computations apply to the input gate i_t and the output gate o_t. Moreover, a new candidate cell state $\tilde{c}_t$ is computed as $\tilde{c}_t = \tanh(W_c y_t + U_c h_{t-1} + b_c)$ and based on the forget gate and the input gate the cell state is updated according to

$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t, \tag{28}$$

where $\circ$ denotes an elementwise product. Here, the roles of the input and the forget gate become clear: the forget gate regulates how much of the previous cell state c_{t-1} is kept, and the input gate determines how much of the new candidate state $\tilde{c}_t$ is added. Finally, the new hidden state h_t is computed as $h_t = o_t \circ \tanh(c_t)$. The essential idea of this process is as follows: In a classical recurrent network where every layer is connected to the previous layer, the backpropagation algorithm to adjust the network's weights yields exploding or vanishing gradients. Thus, the network cannot be trained properly, in particular the hidden layers that are closer to the input layer than to the output layer are not treated in an effective way, compare [32]. As a remedy, Hochreiter and Schmidhuber in [31] proposed a network that carefully regulates the information flow within the network by introducing various gates in the network (input, output, and forget). Figure 1 provides a schematic illustration of a single LSTM unit or cell as described above.

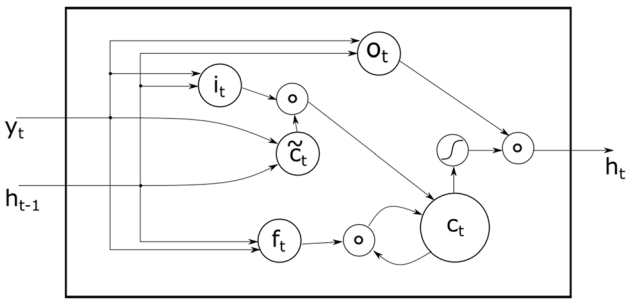

Fig. 1 Schematic illustration of a single LSTM cell. The figure shows the information flow in a cell between the various gates

In many applications, the output of the LSTM network is solely based on the last hidden state h_T. However, in our situation we feed the network with a TAC sequence $y = (y_1, \ldots, y_T)$ and want the output to be the corresponding estimated BrAC sequence $\tilde{u} = (\tilde{u}_1, \ldots, \tilde{u}_T)$. This is known as the sequence-to-sequence setting. The complexity of an LSTM network is mainly governed by the dimensionality of the output h_t. Thus, in order to get scalar values $\tilde{u}_t$ based on h_t, we apply a feed-forward network to the hidden states h_t whose output is the desired estimated BrAC value at time t, $\tilde{u}_t$.

When training such a network, the matrices W_f, W_i, W_o, W_c, U_f, U_i, U_o, and U_c as well as the bias terms b_f, b_i, b_o, and b_c are adjusted to minimize a given loss function. Such a loss function typically relies on a difference between the network output and the true output. Hence, a simple loss function could be given by $\mathscr{L} = \mathrm{MSE}(u, \tilde{u})$, where u denotes the true BrAC signal and $\tilde{u}$ denotes the estimated BrAC (eBrAC) signal, and MSE is the mean squared error between the two signals.

As described in Sect. 1, the available data sets to train the network on are quite small for the alcohol biosensor problem. This lack of data may result in several problems. First, the insufficient amount of data may hinder the model's ability to learn appropriately, i.e., the network does not learn a suitable relationship between TAC and BrAC signals, and second, the model may tend to overfit on the training data set and thus its generalization capabilities can be poor. To avoid these issues, here we propose a physics-informed learning scheme where the a priori physics information derived from the physical model (1)–(6) is used as a regularization term.

Based on the mechanisms developed in [20, 33–36], we can obtain a physics-informed LSTM network by augmenting the loss function with an additional penalty term. Moreover, since the natural input and output of LSTM networks are time series (rather than data points (t, η) in the time-space domain), we can make use of the previously described formulation of the system (1)–(6) as an abstract parabolic system and its discretization in both time and space.

In [37], the authors use the minimization problem

$$u^* = \arg\min_{\tilde{u}} \sum_{k=1}^{T} |\mathscr{y}_k^N(\tilde{u}) - y_i|^2 + ||\tilde{u}||^2 \tag{29}$$

for the input or BrAC estimation. Here, $||\tilde{u}||$ denotes a regularization term with a suitable norm $||\cdot||$ and $\mathscr{y}^N(\tilde{u})$ denotes the forward solution of the physical model, i.e., the eTAC based on the eBrAC using the discretized forward population model as in (24).

In this work, we want to make use of this idea in the context of the LSTM network. To this end, we propose to

augment the loss function of the neural network that is fully based on data by a physics term involving the squared difference between $\mathscr{y}^N(\tilde{u})$ and y, i.e., the TAC output of the eBrAC $\tilde{u}$ under the physical model (24) is encouraged to match the true TAC output y. That way, the network is physics-informed in the sense that physically meaningless eBrAC signals are penalized and this additional penalty shall help the network to learn, although the available training data set is rather small.

More precisely, we define the loss function as

$$\mathscr{L} = \mathrm{MSE}(u, \tilde{u}) + \beta\Big(\mathrm{MSE}(y, \mathscr{y}^N) + \kappa_1 \|\tilde{u}\|_2^2 + \kappa_2 \mathrm{SV}(\tilde{u})\Big), \tag{30}$$

where MSE denotes the mean squared error, $\|\cdot\|_2$ is the Euclidean norm, and $\mathrm{SV}(x) = \sum_i (x_i - x_{i+1})^2$ is the squared variation of x. κ_1 and κ_2 denote tuning parameters, and β is the parameter that governs how much emphasis is placed on the physical information. Assuming that the number of drinking episodes in the training batch is N_b, the loss function can be written as

$$\mathscr{L} = \frac{1}{N_b}\sum_{j=1}^{N_b}\left(\frac{1}{T}\sum_{t=1}^{T}\left((u_{j,t} - \tilde{u}_{j,t})^2 + \beta\left(y_{j,t} - \mathscr{y}_{j,t}^N\right)^2\right) + \beta\kappa_1 \sum_{t=1}^{T}\tilde{u}_{j,t}^2 + \beta\kappa_2 \sum_{t=1}^{T-1}(\tilde{u}_{j,t} - \tilde{u}_{j,t+1})^2\right). \tag{31}$$

Note that the deconvolution of BrAC from TAC based on the minimization problem (29) as used in [37] is not suitable for real-time BrAC estimation. That is, eBrAC signals can change based on new incoming data. Assume a recorded TAC signal $y = (y_1, \ldots, y_T, y_{T+1})$. Then, the two sequences $(u^*(y_1, \ldots, y_T))_{t=1}^T$ and $(u^*(y_1, \ldots, y_T, y_{T+1}))_{t=1}^T$ are not necessarily identical. Our proposed physics-informed LSTM network does not possess this property; the eBrAC signal $(\tilde{u})_{t=1}^T$ is only based on $y_1, \ldots, y_T$. This enables our approach to yield eBrAC signals in real time, a feature that is especially relevant for the use of BrAC estimation in wearables.

4 Network architecture

As mentioned in Sect. 1, TAC is not only a function of BrAC. Instead, the expression from BrAC to TAC depends on a variety of other covariates as well. These can roughly be grouped into body covariates, drinking episode level covariates and behavioral covariates. Body covariates include, but are not limited to, gender, age, body mass index, ethnicity, and resting metabolism. On the drinking episode level, the alcohol dosing pattern, ambient temperature, blood pressure, and heart rate might influence the functional relationship between BrAC and TAC. Finally, behavioral covariates such as the past 28 or 90 day drinking behavior can lead to adaptions in the body and hence affect the diffusion process as well. Accounting for all these covariates in the model might explain the observed variance in the TAC expression and eventually lead to improved BrAC estimates.

When handling the various covariates, the issue arises that some covariates depend on time, such as TAC, heart rate, and ambient temperature, while others are constant over time, such as the age or the gender of the human subject. Consequently, these covariates should be treated differently within the neural network. Here, we propose a network architecture consisting of three separate networks. Network 1 is a physics-informed LSTM network as described above to process the time-dependent covariates. Network 2 is a standard feed-forward network that processes the time-independent covariates. Finally, in network 3 the outputs of these networks are combined in a feed-forward network to yield the eBrAC. A schematic diagram of the three networks is shown in Fig. 2. To be more precise, the input to network 1 is time series that contain equally spaced measurements of TAC. In some of our numerical studies, we also process heart rate time series in this network, see Sect. 6. In that case, the processed time series are two-dimensional. The outputs of network 1 are again time series, and we choose the dimension of the output space to be 50. At the same time, network 2 processes 25 time-independent covariates per drinking episode in a fully connected feed-forward network (see "Appendix A" for a list of the covariates). The output dimension of this network is 5. Then, at every time step, we concatenate the outputs of network 1 and network 2 and pass the resulting vector of dimension 55 to another fully connected

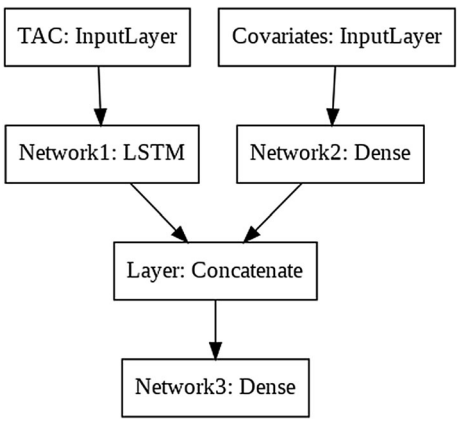

Fig. 2 Structure of the proposed network for the handling of the time-dependent covariates (grouped as "TAC") and the time-independent covariates

network, network 3, whose output is a one-dimensional time series, the eBrAC signal.

5 Uncertainty quantification using dropout

Given the observed large variance in BrAC/TAC data pairs, a completely deterministic approach to BrAC estimation does not appear to be very promising. Therefore, to account for the variability and uncertainty not explained by the model, a quantitative assessment of the uncertainty is key to meaningful estimates. Previous works [37, 38], relying fully on the physical model (1)–(6), assumed the parameters $q \in Q$ to be random and estimated their distribution. Based on this distribution, values for q were sampled to produce credible bands for the eBrAC. Although the present work also uses the model (1)–(6), we do not aim at estimating the distribution of q, but we aim at fitting the network parameters ω where $\omega = \{W\}$ with $W = (W_f, W_i, W_o, W_c, U_f, U_i, U_o, U_c, b_f, b_i, b_o, b_c)$ is the set of all the network parameters to fit. In this context, an uncertainty quantification could be based on the idea of a Bayesian neural network. In Bayesian neural networks, the network parameters ω are assumed to be random following some prior distribution. The goal is then to come up with a posterior distribution for ω based on the data. The primary caveat of a Bayesian network is the complexity of these models, so that more sophisticated models quickly become intractable. As a remedy, a common method is to consider an approximation to the full posterior distribution, a process that is known as variational inference [39]. Only recently, a remarkable connection between this variational inference and the common regularization technique known as Monte-Carlo dropout was found by Gal in [40]. Here, we will briefly describe the uncertainty quantification that is derived from this relation, following the steps taken in [41].

To this end, we consider a simple problem where the input y and output u of the network are scalar. Later, we will make the appropriate adaptions for the time series network considered here. We consider the network parameters ω to be random and we assume they have some prior distribution $p(\omega)$. Further, we need to define a likelihood distribution $p(u|y, \omega)$, i.e., the output distribution given the observed variables y and the network parameters ω. Such a likelihood can be taken to be a Gaussian as

$$p(u|y, \omega) \sim \mathcal{N}(u; f^{\omega}(y), \tau), \tag{32}$$

where f^{ω} denotes the parameterized function that relates the input y and the network output u by $u = f^{\omega}(y)$. In our case, f will be a physics-informed LSTM network as described above. Moreover, τ denotes the variance or noise.

Let Y and U be given datasets. Then, using Bayes' theorem, obtain the posterior distribution of ω as

$$p(\omega|Y, U) = \frac{p(U|Y, \omega)p(\omega)}{p(U|Y)}. \tag{33}$$

The denominator of this expression, $p(U|Y)$ is known as the model evidence. Written out, it takes the form of an integral

$$p(Y|X) = \int p(Y|X, \omega)p(\omega)d\omega. \tag{34}$$

This integration is called marginalization and it is well known that the exact marginalization often becomes unfeasible for more complex problems. This is also the case for the LSTM network considered here. Hence, we choose to look for a variational approximation q_θ to approximate the posterior distribution as $q_\theta(\omega) \approx p(\omega|Y, U)$. Here, a distribution q, parameterized by θ, is chosen such that it is easy to evaluate. A comprehensive text on the topic of variational approximation can be found in [39]. Ideally, we want the approximation $q_\theta(\omega)$ to be as close to the true posterior distribution $p(\omega|Y, U)$ as possible. To achieve this, our goal is to minimize the Kullback–Leibler divergence between the two distributions with respect to the parameter θ. This divergence is given by

$$\mathrm{KL}(q_\theta(\omega)||p(\omega|Y, U)) = \int q_\theta(\omega) \frac{q_\theta(\omega)}{p(\omega|Y, U)} d\omega. \tag{35}$$

Minimizing the Kullback–Leibler divergence is equivalent to maximizing the so called evidence lower bound (ELBO) with respect to the parameterization θ. This ELBO is defined as

$$\int q_\theta(\omega) \log p(U|Y, \omega)d\omega - \mathrm{KL}(q_\theta(\omega)||p(\omega)). \tag{36}$$

The described method is known as variational inference and is well established in the field of Bayesian modeling. Written out, the ELBO takes the form

$$\sum_{i=1}^{N} \int q_\theta(\omega) \log p(u_i|f^{\omega}(y_i))d\omega - \mathrm{KL}(q_\theta(\omega)||p(\omega)), \tag{37}$$

where N denotes the number of data pairs (y_i, u_i). Now, let us take a more specific look at the time series problem considered here. In this case, the inputs take the form $y_i = (y_{i,1}, y_{i,2}, \ldots, y_{i,T})$ and the outputs take the form $\tilde{u}_i = (\tilde{u}_{i,1}, \tilde{u}_{i,2}, \ldots, \tilde{u}_{i,T})$. For simplicity, we drop the index i for now and focus on the time index t. Then, the LSTM network described in Sect. 3 defines functions f_h^{ω} that relate the hidden states h_t with the input arguments as $h_t = f_h^{\omega}(y_t, h_{t-1})$ for $t = 1, \ldots, T$. Further, the output $\tilde{u}_t$ depends

on the hidden state h_t of the LSTM network and the output k of the time-independent covariate network via $\tilde{u}_t = f_u^{\omega}(h_t, k)$. In that way, the full output sequence $\tilde{u}$ can be expressed as

$$\begin{aligned}\tilde{u} &= f_u^{\omega}(h_0, \ldots, h_T, y_1, \ldots, y_T, k) \\ &= f_u^{\omega}\big(h_0, f_h^{\omega}(y_1, h_0), \ldots, f_h^{\omega} \\ &\quad \big(y_T, f_h^{\omega}\big(y_{T-1}, f_h^{\omega}\big(\ldots f_h^{\omega}(y_1, h_0)\big)\big)\big)\big), y_1, \ldots, y_T, k\big)\end{aligned} \tag{38}$$

where $h_0 = 0$.

Using this notation, the objective from (37) takes the form

$$\begin{aligned}&\sum_{i=1}^{N} \int q_\theta(\omega) \log p(y_i | f_u^{\omega}(h_0, \ldots, h_T, x_1, \ldots, x_T, k)) d\omega \\ &\quad - \mathrm{KL}(q_\theta(\omega) || p(\omega)).\end{aligned} \tag{39}$$

Then, to approximate the integral, we use Monte-Carlo integration and obtain the objective

$$\begin{aligned}&\sum_{i=1}^{N} \log p(y_i | f_y^{\hat{\omega}}(h_0, \ldots, h_T, x_1, \ldots, x_T, k)) \\ &\quad - \mathrm{KL}(q_\theta(\omega) || p(\omega)),\end{aligned} \tag{40}$$

that is based on a single sample $\hat{\omega}$ from the variational distribution $q(\omega)$. It is known that (40) is an unbiased estimator for (39). For more information on Monte-Carlo integration we refer to [42].

Finally, for the parameterization of $q_\theta(\omega)$, we follow the approach of Hinton/van Camp [43] to factorize over the weight matrices W_j of the network as

$$q_\theta(\omega) = \prod_j q_{M_j}(W_j) \tag{41}$$

where $M_j = \mathrm{mean}(W_j)$ denotes the mean weight matrix and $q_{M_j}(W_j)$ is given by $q_{M_j}(W_j) = M_j \cdot Z$ where Z is a diagonal matrix whose elements $z_{k,k}$ are i.i.d. Bernoulli distributed as $z_{k,k} \sim \mathrm{Bernoulli}(p)$ with $p \in [0, 1]$ to be defined.

This choice for the variational distribution immediately gives rise to the mentioned technique of dropout [44]. Dropout is a common stochastic regularization technique, usually used to prevent a neural network from overfitting. Using dropout, in each training iteration, a randomly selected set of neurons in the network is deactivated. The main idea behind this is that the ever-changing structure of the network hinders it to just learn and reproduce the training data. Also, since the network architecture is different on every training batch, one can consider the final outcome as an average over an ensemble of slightly different neural networks. For a more detailed description of dropout regularization, we refer the reader to [44]. Usually, a certain percentage p is chosen and then a p fraction of all the neurons is masked randomly in every training step. This description is equivalent to randomly setting columns of the weight matrices M_j to zero, which in turn is equivalent to the multiplication of M_j with the random diagonal matrix Z whose entries are Bernoulli(p) distributed.

In [40, 41], it is shown that finding the optimal network parameters ω in a neural network using dropout is essentially the same objective as the variational inference objective in (37) with the special choice of variational distribution $q_\theta(\omega)$ as in (41). This implies that every neural network trained with dropout can be considered as a special type of Bayesian neural network. It is this observation that enables us to obtain an uncertainty quantification in the sense of a Bayesian neural network within a feasible computational time. In practice, this remarkable connection means that a neural network, trained using dropout, can be used to assess the uncertainty of the outcomes in the inference phase by simply using dropout in the inference phase as well. Hence, by sampling from the Bernoulli random matrix Z we automatically obtain samples of the approximate posterior $q(\omega)$ that can be used to generate samples of the eBrAC $\tilde{u}$. Thus, based on samples $\hat{\omega}_k$, $k = 1, \ldots, M$, we get a means to produce samples of the eBrAC sequences $\tilde{u}_k = f_u^{\hat{\omega}_k}(y_1, \ldots, y_T)$, $k = 1, \ldots, M$. It is then straightforward to obtain quantities like the mean eBrAC signal as

$$\bar{u} = \frac{1}{M} \sum_{k=1}^{M} f_u^{\hat{\omega}_k}(y_1, \ldots, y_T), \tag{42}$$

as well as credible regions that contain a certain percentage of all the observed samples. In Sect. 6, we will demonstrate this uncertainty quantification technique using data from actual drinking episodes.

6 Numerical studies

6.1 Data

The data that have been used for our numerical studies consist of BrAC, TAC, and covariate data from a sample of 40 participants (50% female, ages 21–33 years, 35% BMI $\geq$ 25.0), who each completed four laboratory drinking sessions. Each participant consumed the same amount of alcohol in all four sessions, which was designed to reach a maximum BrAC peak of .080 (the legal limit for intoxication in the US) for that person based on body water weight. There were three patterns of consumption: participants consumed the alcohol in 15 min (single dose) in two of the sessions, in two 15 min periods spaced 30 min apart (Dual Dose) in one session, and over 60 min (Steady Dose) in one session. An Intoximeter IV breath analyzer

(Intoximeters, Inc., St. Louis, MO) was used to obtain BrAC, and two SCRAM-CAM devices (Alcohol Monitoring Systems, AMS, Littleton, CO) were placed on the upper arms to obtain TAC. Measurements were taken at 10 min intervals over the entire BrAC curve. After BrAC returned to .000, the TAC devices continued to record automatically at 30 min intervals until TAC also returned to .000. Several sets of BrAC-TAC data were excluded from analyses due to a change in the protocol, and with two sets of TAC data recorded in each session, our final sample consisted of 256 BrAC-TAC datasets [45]. We also obtained covariates that may affect the BrAC-TAC relationship, including person-level covariates of sex, age, race/ethnicity, body measurements (e.g., hip and waist measurements, percent body fat), and past 90-day drinking behavior and session-level covariates of alcohol dosing pattern and blood pressure readings. A complete list of the covariates available to us is given in "Appendix A".

6.2 Fitting the population model

Before the proposed physics-informed LSTM model can be trained, we first need to find the parameters for the forward population model as in Sect. 2. Here, we use a two-step process. In the first step, we fit the raw parameters $\boldsymbol{q} = [q_1, q_2, q_3, q_4]$ to the individual drinking episodes from the BrAC-TAC dataset using techniques for solving inverse problems for distributed parameter systems that can be found in [11, 15, 17, 26, 46]. Then, in the second step, we remove those episodes for which any of the fitted parameters were larger than 1.5 times the right hand boundary of the third quartile for that parameter, and use the remaining parameter values for $\boldsymbol{q}$ to fit scaled Gamma distributions to q_1, q_2, q_3 and q_4 using the method of moments [47]. In this, we based the scaling on the n th order statistic $\gamma_{i,(n)}$ for $i = 1, 2, 3, 4$. Note that our theory does not require the independence of q_i, $i = 1, \ldots, 4$. However, for simplicity we assume independence in this treatment here. Also note that the choice of a Gamma distribution is not a necessity, but the Gamma distribution is flexible enough to appropriately model the distribution of the parameters, see below. Using this process, we obtained the distributions $q_1 \sim \gamma_{1,(n)}\gamma_1$ with $\gamma_{1,(n)} = 3.64$ and $\gamma_1 \sim \text{Gamma}(3.31, 0.12)$, $q_2 \sim \gamma_{2,(n)}\gamma_2$ with $\gamma_{2,(n)} = 5.99$ and $\gamma_2 \sim \text{Gamma}(4.60, 0.095)$, $q_3 \sim \gamma_{3,(n)}\gamma_3$ with $\gamma_{3,(n)} = 3.49$ and $\gamma_3 \sim \text{Gamma}(2.21, 0.14)$, and $q_4 \sim \gamma_{4,(n)}\gamma_4$ with $\gamma_{4,(n)} = 4.70$ and $\gamma_4 \sim \text{Gamma}(5.80, 0.074)$. Figure 3 shows histograms of the individually fitted parameters q_i and the distribution fit for $i = 1, 2, 3, 4$.

6.3 Numerical results

We split the 256 sets of BrAC-TAC data into a training set containing 192 episodes and a test set with the remaining 64 episodes, with all of each participant's episodes placed either in the training or in the test set to reduce the effects of non-independence of repeated drinking episodes within participants. If not stated otherwise, the only input to the LSTM network is the TAC data, the time-independent covariates are handled according to Sect. 4. All results displayed below show predictions on the test set using an LSTM network that has been trained solely on the training set. The physics-informed LSTM network was implemented in Python, using the deep learning application programming interface Keras. To train the model, we used the Adam optimizer [48]. For the choice of the parameters in the Gamma distributions for $q_1, \ldots, q_4$ we used the values found in the previous subsection.

The particular episodes used to demonstrate our approach were selected because they were generally representative of the particular aspect of our method we were intending to illustrate. Figure 4 depicts four drinking episodes from the test set. It shows the true BrAC signal u as well as the eBrAC signal $\tilde{u}$ which is the mean of 1000 sampled eBrAC signals using the method outlined in Sect. 5 together with 95% credible regions for the BrAC signal. Overall, we see that the mean eBrAC signal quite nicely matches the BrAC signal. The magnitude of the peak is not accurately captured in some cases, but this comes as no surprise, given the observed variation in the data. Moreover, the shown credible regions for the BrAC signal appear to accurately capture the uncertainty of the prediction: the true BrAC signal is within these bounds most of the time.

6.4 Error evaluation

When evaluating the goodness of fit of the model, four quantities are of particular interest. These are the relative L_2 error (rL_2), the relative peak error which is the relative distance between the peak of the true BrAC signal and the eBrAC signal (rP), the relative error of the area under the curve ($rAUC$), and the absolute difference between the peak times of the BrAC signal and the eBrAC signal ($|\Delta T|$).

Table 1 shows these error values for the trained model on both the training and the test set. As can be seen, the errors on training set and test set are very similar. This indicates that the model generalizes fairly well and does not overfit.

The proposed framework allows for the time-dependent input data to be vector valued as well. Hence, it is easy to

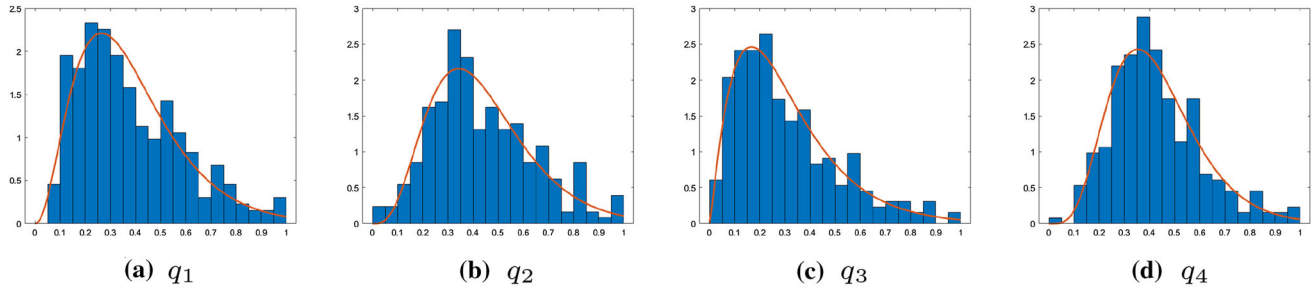

Fig. 3 Scaled histograms of individually fitted q_i parameters together with probability density functions of distributional fits

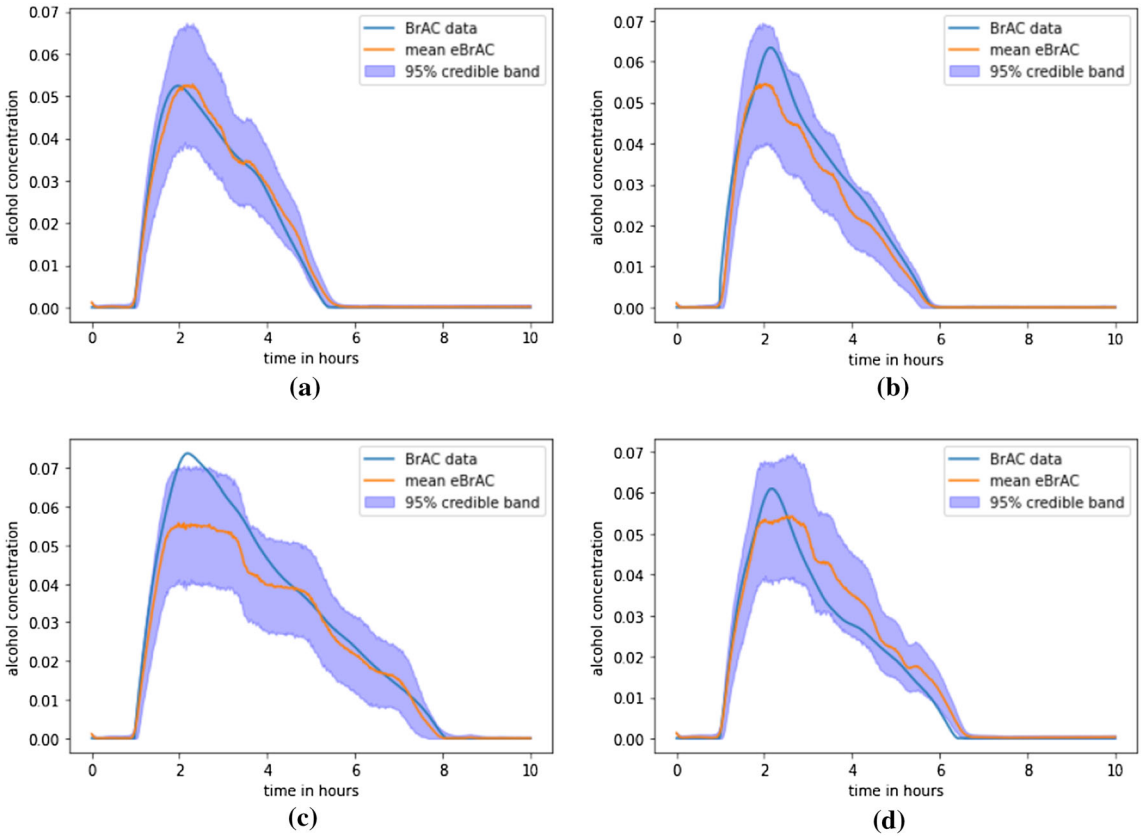

Fig. 4 Performance of the physics-informed LSTM model on four drinking sessions from the test set. The figures show the BrAC data u, the predicted mean eBrAC signal $\bar{u}$, and credible regions for the BrAC signal

Table 1 Error values on training and test set

	rL_2	rP	$rAUC$	$\lvert\Delta T\rvert$
Training	0.2168	0.1650	0.1496	0.2186
Test	0.2183	0.1565	0.1590	0.2547

Table 2 Error values on training and test set, heart rate included

	rL_2	rP	$rAUC$	$\lvert\Delta T\rvert$
Training	0.2065	0.1536	0.1287	0.2339
Test	0.2295	0.1491	0.1682	0.2538

incorporate multiple time-dependent covariates into the model. In Table 2, we show the resulting error values when TAC and the heart rate are used as input signals for the LSTM network. As can be seen, the model including the heart rate performs slightly better on the training set, but no clear statement can be made for the test set. It appears that the heart rate does not improve the performance or the

insight of the model significantly. More on the importance of covariates in Sect. 6.8.

6.5 Impact of the parameter β

As described in Sect. 3, we use the parameter β to balance the data and the physics information in the loss function (30). Intuitively, a lower β yields an estimate that aligns best with the training data, while a larger β tries to match the computed TAC signal of the estimate, $\mathcal{y}^N(\tilde{u})$, with the true TAC signal y. Figures 5 and 6 show the numerical results for varying values of β based on two drinking sessions from the test set. In either figure, we display results for $\beta \in \{0.01, 0.05, 0.1, 0.25\}$. In Fig. 5, the signals u and y are roughly of the same magnitude. For $\beta = 0.01$, the agreement between y and $\mathcal{y}^N(\tilde{u})$ is rather poor and also $\tilde{u}$ does not quite match the signal u, see Fig. 5a. For $\beta = 0.05$ (Fig. 5b), the regularization mechanism yields a TAC signal $\mathcal{y}^N(\tilde{u})$ that better matches y. Also the agreement between u and $\tilde{u}$ is good. For even larger values of β the agreement between $\mathcal{y}^N(\tilde{u})$ and y becomes significantly better, compare Fig. 5c and d. However, the estimated signal $\tilde{u}$ deteriorates (see Fig. 5c), and even becomes unphysical as the negative values in Fig. 5d show. This indicates that indeed an intermediate value of β yields the best balance between data-fitting and physical knowledge. How important the physics information actually is can be seen in Fig. 6. Here, the TAC signal y is of much smaller magnitude than the BrAC signal u. Since situations like these are rare in the training data, the model with little to no physics information yields an estimated signal $\tilde{u}$ that is close to zero, compare Fig. 6a. Then, as β increases, the signal $\tilde{u}$ clearly shows a drinking episode, see Fig. 6b. Note that also in this case higher values for β lead to a non-physical behavior, as Fig. 6c and d show. Note further that although the signal $\tilde{u}$ in Fig. 6b does not match the true signal u, the result is much better than in the non-physics-informed case where there appears to be no drinking at all. Hence, the physics-informed mechanism proves especially useful in situations that are not or barely covered in the training data. Thus, this mechanism is important for the extrapolation or generalization capabilities of the model.

When looking at Figs. 5 and 6, the eBrAC signal becomes less smooth as the regularization parameter β increases, i.e., it starts to oscillate. This is somewhat counterintuitive, as a regularization usually enhances the regularity. However, in the context of this work, the physics-informed regularization is not meant to make the

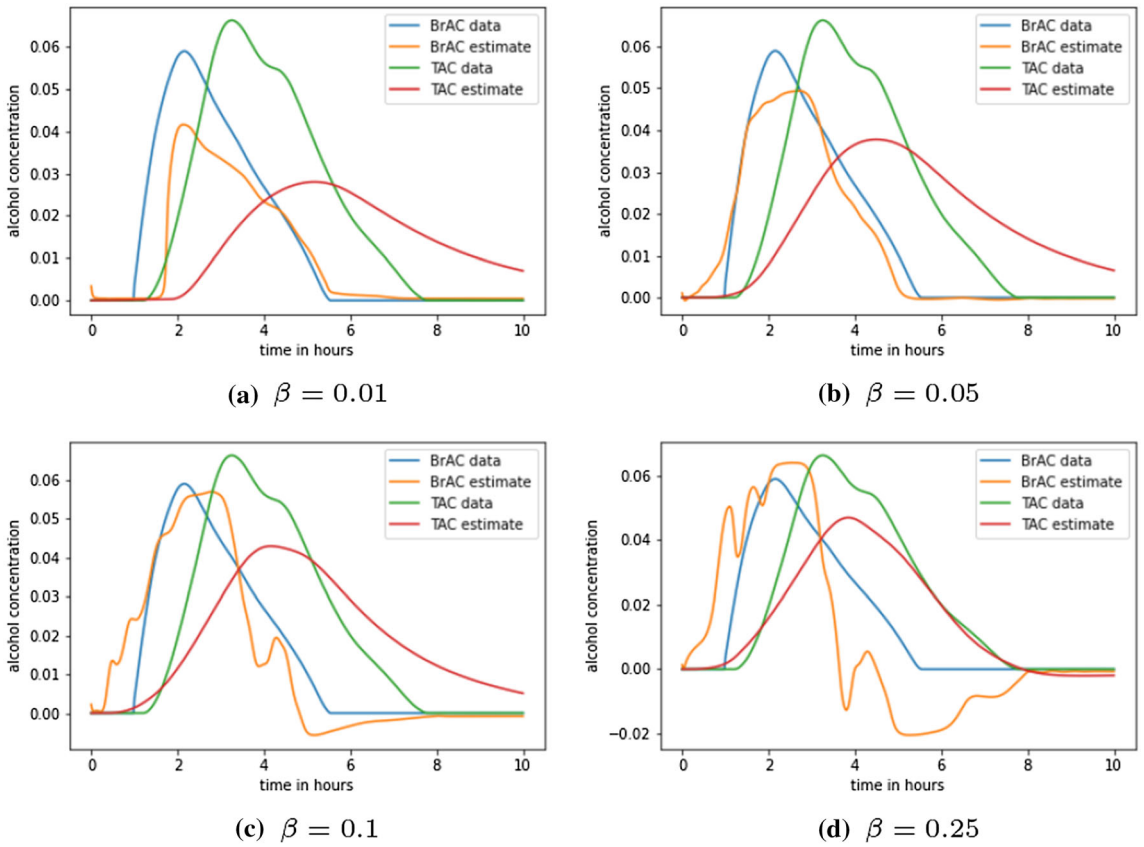

Fig. 5 Results of the BrAC estimation using different values for the parameter β on a single drinking episode

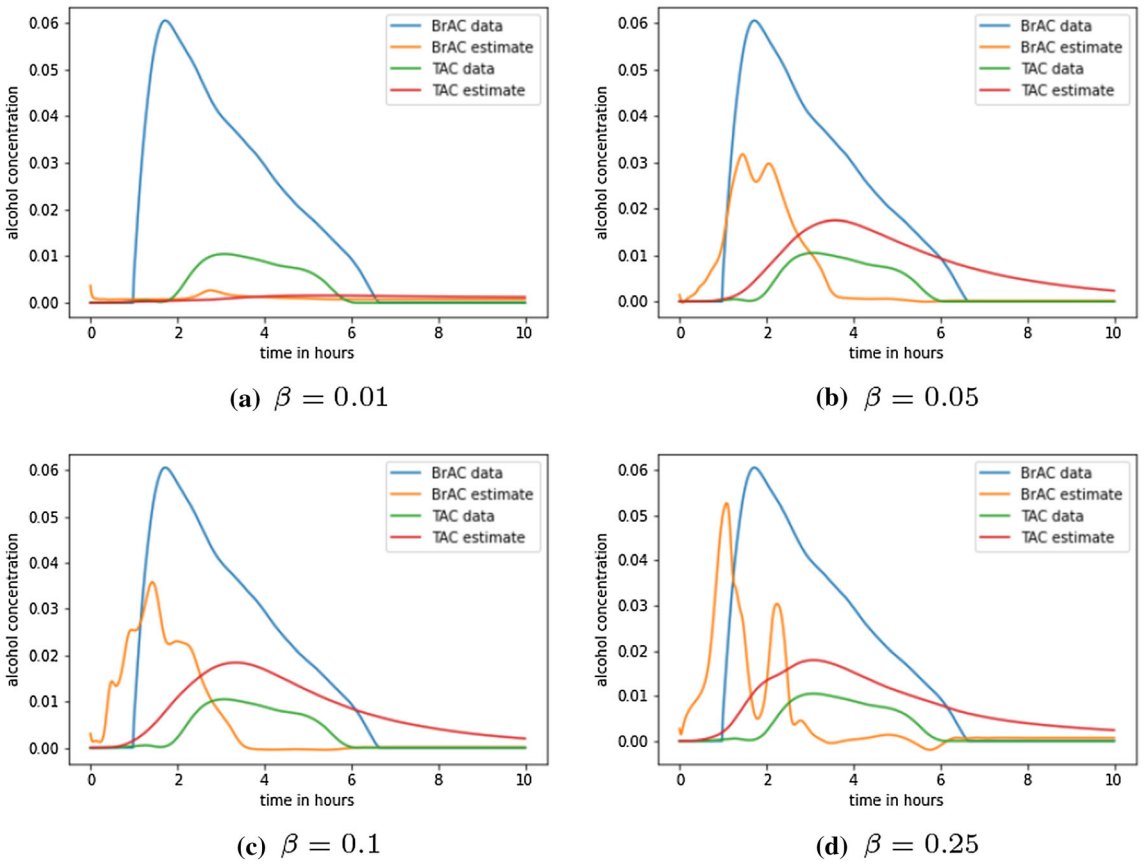

Fig. 6 Results of the BrAC estimation using different values for the parameter β on a single drinking episode

eBrAC signal smoother, but to make it physically meaningful. In that sense, the observation is not contradictory. Note that the loss function in (30) also contains terms related to the variation of the eBrAC signal. By fine tuning the parameters κ_1 and κ_2 in (30) a higher smoothness of the eBrAC signals could be obtained. We do not discuss this matter in the proof of concept reported here.

6.6 Role of the bias terms

The LSTM network described in Sect. 3 includes bias terms b_f, b_i, b_o in the forget, input, and output gate, respectively. These bias terms can be used to encode a general behavior of drinking sessions that is present in most episodes from the training set. A problem, however, arises if the training set is rather small or does not cover the whole range of realistic samples, as in the context of this alcohol biosensor problem. Then, features present in training drinking episodes can mistakenly be generalized to be features of every drinking episode. This results in a tradeoff: strong bias terms yield good results on the training data set, but might not be applicable to data coming from outside. Without bias terms, on the other hand, the estimated signals might be missing important information. Figure 7 demonstrates this behavior. In Fig. 7a, the eBrAC signals obtained with or without bias terms are very similar. The only difference is that the estimate with bias terms correctly predicts the start of the drinking episode, whereas the estimate without bias does not. The reason for this behavior is the delay between the TAC and the BrAC signal. In Fig. 7b, both estimates predict the start of the episode correctly, but the estimate without bias terms shows some non-physical behavior. Figure 7c, however, shows an example where the estimate with bias is far off the true signal, whereas the estimate without bias captures the true signal much better. We can attribute this behavior to a TAC signal that has a lower magnitude than most TAC signals in the training data. These observations suggest that an LSTM model without the bias terms will be advantageous for the application to a wide variety of different drinking episodes.

6.7 Results on field data

So far, the BrAC and TAC for all the presented results were collected in the laboratory under controlled conditions following a prescribed dosing protocol. Now, we want to use our approach for the estimation of BrAC under

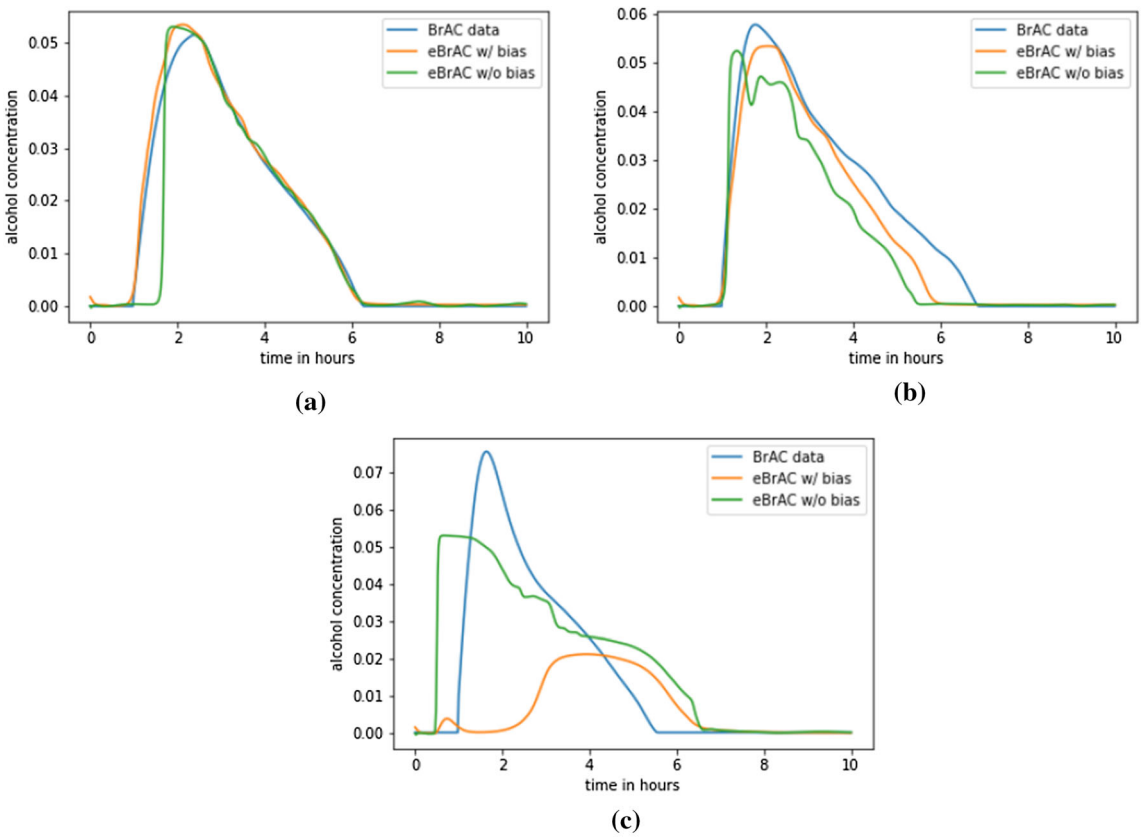

Fig. 7 Results of the BrAC estimation using models with and without bias terms. We display three drinking episodes from the test set

naturalistic drinking conditions in the field. In [19], it was shown that the direct application of a BrAC-estimating model to field data can be difficult, especially if the naturalistic drinking episodes differ a lot from the training drinking episodes with regards to shape and magnitude. The three naturalistic drinking episodes considered here were collected by one of the authors (S.E.L.) using the WrisTAS 7^{TM} transdermal alcohol biosensor (Giner, Inc., Waltham, Massachusetts). We trained a non-physics-informed LSTM model (labeled LSTM) and a physics-informed model (labeled PILSTM) with the training set described earlier, containing only laboratory controlled drinking episodes. Figure 8 shows that the overall performance of the models on the naturalistic drinking data is worse compared to the laboratory controlled drinking sessions. However, the PILSTM yields results that still capture the essential dynamics of the true signals with respect to shape and magnitude, whereas the default LSTM shows erratic behavior. This can especially be seen in Fig. 8a and b where the non-physics-informed estimates exhibit completely unphysical artifacts toward the end of the episodes. The use of the physics regularization effectively suppresses these artifacts. Even episodes without these artifacts display significant oscillation, see Fig. 8b. We believe that this can be attributed to the fact that the TAC signals from the field episodes are quite different than those of the training set. Hence, the model could not appropriately learn these situations. Of course, a possible remedy would be the use of a richer training set that includes a variety of naturalistic drinking episodes. Unfortunately, as described earlier, data gathering for the problem at hand is expensive and time-consuming. Thus, the physics-informed mechanism appears to be valuable in that it mitigates this lack of training data.

6.8 Importance of covariates using shapley values

As outlined in Sect. 1, the proposed BrAC estimation model is one of the first to include a range of covariates, such as body covariates, session level covariates, and behavioral covariates. Hence, thus far, only little is known about the importance of these covariates for the BrAC estimation and how they affect the TAC expression. Here, we want to qualitatively and quantitatively assess the importance of the covariates. To this end, we employ an idea from cooperative game theory, the so-called Shapley values [49]. Named after Lloyd Shapley, Shapley values

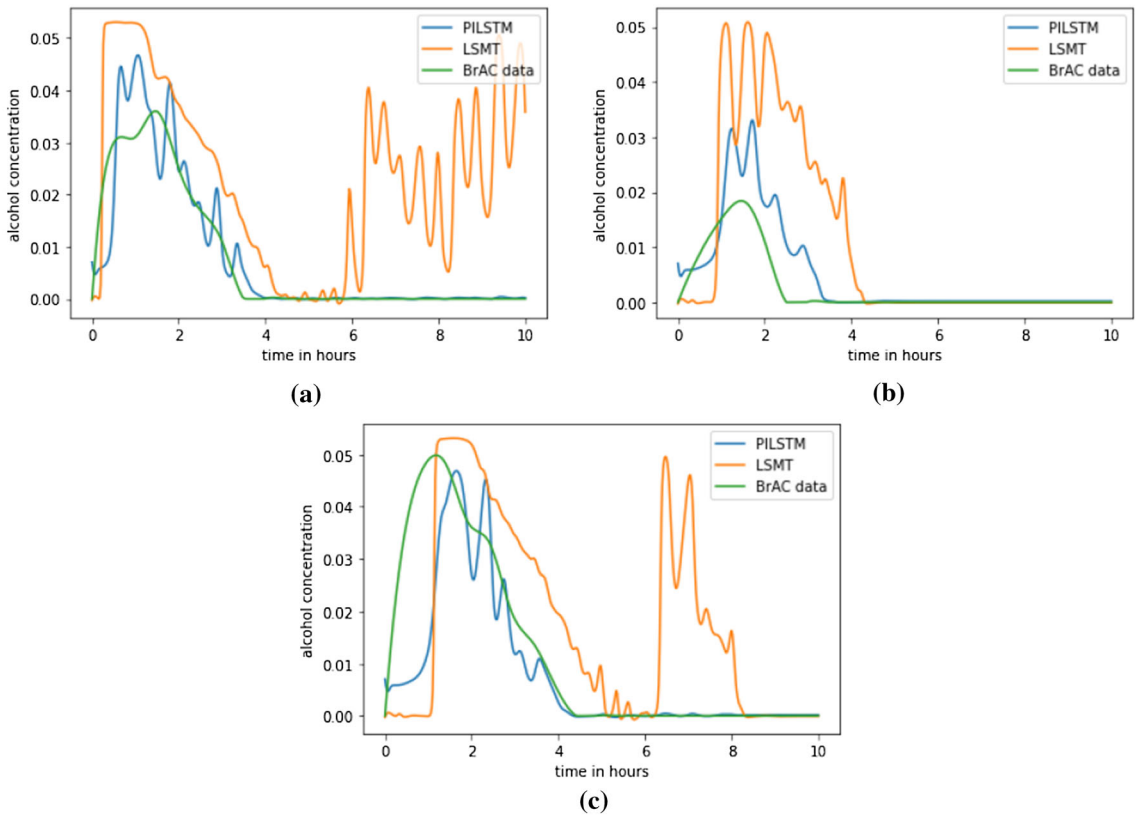

Fig. 8 Application of the LSTM model to field data drinking sessions. We show the BrAC data, eBrAC signals obtained from the LSTM model without physics (labeled LSTM), and eBrAC signals obtained from the physics-informed LSTM model (labeled PILSTM)

can be applied to machine learning problems to find each covariate's contribution to the result.

Formally, Shapley values are defined as follows. Assume a set D of d players together with a function $w : 2^D \to \mathbb{R}$ such that $w(\emptyset) = 0$. The function w maps every subset S of players to the so-called worth of coalition, i.e., the reward that the players in S can obtain cooperatively. Now, in the setting of machine learning, the players can be thought of as the covariates and the worth of coalition is the output of the network when the covariates in S were used as input. Then, based on this definition, Shapley values assign to every single player i its contribution $\phi_i(w)$ according to the formula

$$\begin{aligned}\phi_i(w) &= \sum_{S \subset D \setminus \{i\}} \frac{\mathsf{S}!(d-\mathsf{S}-1)!}{d!}\left(w(S \cup \{i\}) - w(S)\right) \\ &= \frac{1}{d}\sum_{S \subset D \setminus \{i\}} \binom{d-1}{\mathsf{S}}^{-1}\left(w(S \cup \{i\}) - w(S)\right).\end{aligned} \tag{43}$$

well as the operator The essential idea behind this is as follows. For every possible coalition S that does not include player i, the gain of adding i to that coalition, i.e., $(w(S \cup \{i\}) - w(S))$ is computed. Then, this contribution of player i is averaged over all possible coalitions. To apply this characterization to our problem, we need to come up with a suitable function w. We choose $w(S)$ to be the area under the curve (AUC) of the eBrAC signal when the covariates in S have been used for prediction.

Figure 9 displays the importance and influence of a selection of the available covariates using Shapley values on our whole data set.

When interpreting these results, one should keep in mind that the data basis is fairly small. Hence, outlier values in the data set may distort the results. Moreover, the chosen function w only measures effects that are averaged over the drinking episode. If some covariates affect the eBrAC signal at a specific point during the episode, the function w would not be able to measure this accurately. We note that these are serious limitations of this approach. Figure 9 shows that the body covariates such as body mass index, resting metabolism, skeletal muscle, age, and some blood pressure characteristics are among the most important covariates. The picture drawn here is that younger, physically more active persons with a lower BMI have a higher predicted BrAC AUC than older, physically less active persons with a higher BMI. Unfortunately, only little is known about the biology of BrAC to TAC expression. It

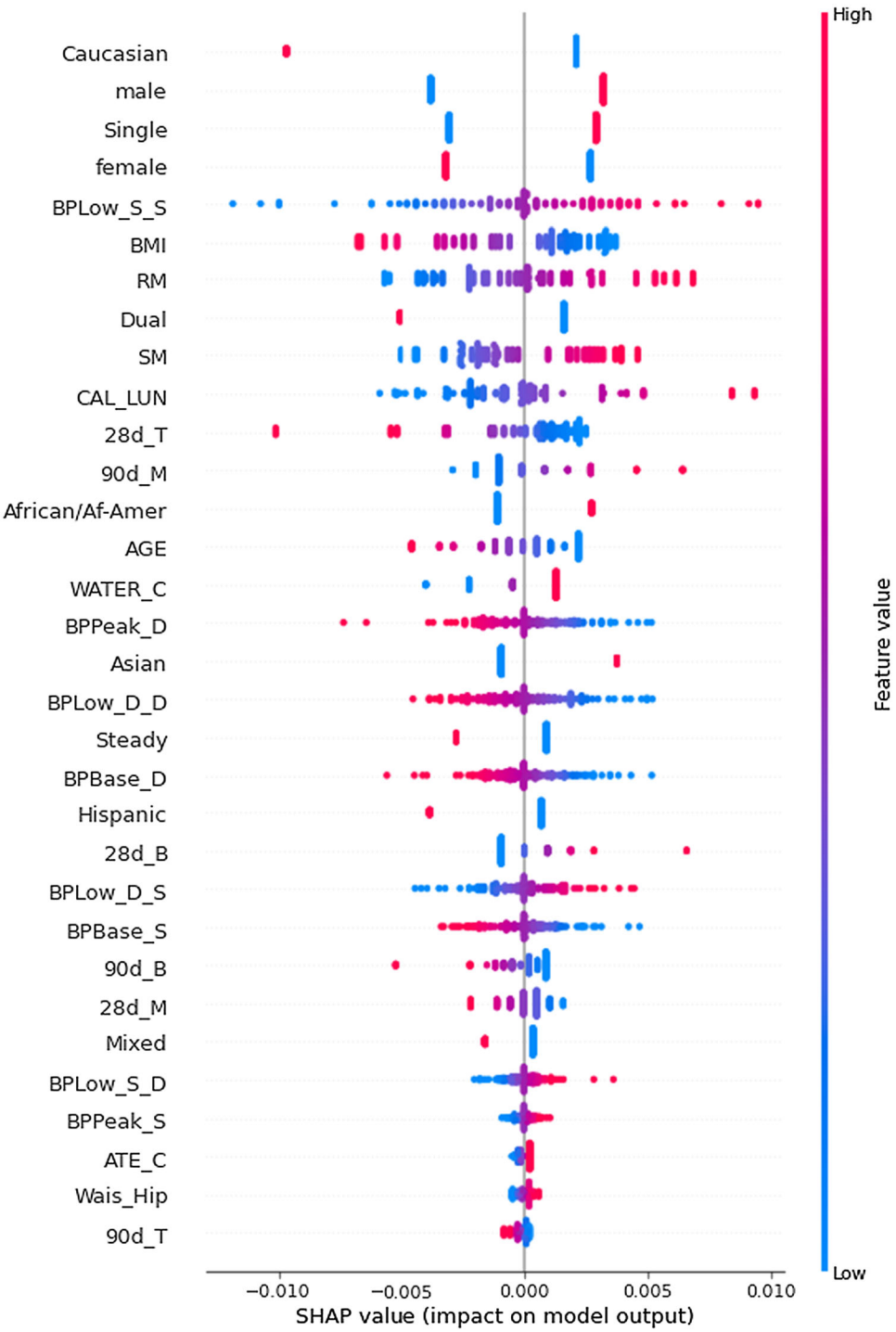

Fig. 9 Shapley values of a selection of covariates. For a description of the covariates, refer to "Appendix A". The order of the covariates is based on the Shapley value importance. Longer bars indicate more data points in that region, while points indicate single data points. The color encoding can be seen as a heat map: high covariate values are red, low values are blue. For 0/1 covariates such as "female" red means "yes" and blue means "no." The impact of a covariate on the BrAC AUC is indicated by the distance from the grey middle bar

is hence difficult to check these observations for plausibility.

A typical BrAC signal has an AUC value on the order of ≈ 0.15. Based on the obtained results, even the most important covariates have an impact of only 0.01 on the estimated signal. This suggests that while the covariates are important to consider in the model, they do not significantly affect the result. Note that we observe that the participants' gender appears to be of importance with a clear distinction between male and female subjects. For male participants, the BrAC estimations have a larger AUC than for females. However, we also observe a difference in the baseline BrAC AUC between males and females of about 0.029. It is thus unclear whether the observed effect can really be attributed to the gender or whether the underlying calibrated alcohol dosing of the experiment led

to these differences. This demonstrates the difficulty in interpreting Shapley values correctly. Similar observations can be made for the ethnicity and dosing pattern. Further research is required.

7 Discussion and remarks

In this work, we considered a physics-informed LSTM model for the estimation of breath alcohol based on transdermal alcohol. We proposed a network architecture that takes into account both time-dependent and time-independent covariates, and we showed how the information encoded in a PDE model can be made readily accessible to a time series network using linear semigroup theory. In addition, we demonstrated how the dropout regularization technique can be employed to yield approximations to credible bands for the estimated signals. In a rather extensive numerical study, we presented results on an actual laboratory collected data set, and we investigated the interplay between the data-driven terms and the physics-driven terms in the loss function.

This study was motivated by the desire to improve earlier approaches and methodologies for the estimation of BrAC from biosensor TAC measurements using recent advances in time series machine learning. The proof-of-concept presented in this work is very encouraging and shows advantages in two particular aspects. First, the flexibility of the underlying neural network models allows for the incorporation of covariate information that has not been taken into account so far. Second, the hybrid combination of a data-driven approach with a physical model appears to be especially valuable in the scarce data setting that is a fundamental characteristic of this application.

Further, we compare the performance of the LSTM model reported on here with two other machine learning models for the TAC-based BrAC estimation. One of these models is a generative adversarial network-based (GAN) approach that can be found in [20], and the other one is a covariate-dependent hidden Markov model (HMM) [19]. In Table 3, we assemble the average error values of the three models for comparison. As in Table 1, we display the average relative L_2 error, the average relative peak error, the average relative area under the curve (AUC) error, and the average absolute peak time error. All error values have been computed on the test set. Based on this comparison, the LSTM model appears very competitive: Compared to the HMM, it reduces the relative L_2 error by 58%, the relative peak error by 33%, the relative AUC error by 48%, and the time to peak error by 35%. Compared to the GAN model, this reduction is even 72% (rL_2), 76% (rP), 75% ($rAUC$), 62% ($|\Delta T|$). A more in-depth controlled comparative study of these models is in progress.

Table 3 Comparison of error values on test set

	rL_2	rP	$rAUC$	$\lvert\Delta T\rvert$
GAN	0.7761	0.6713	0.6467	0.6764
HMM	0.5174	0.2349	0.3091	0.3958
LSTM	0.2183	0.1565	0.1590	0.2547

Looking ahead, we intend to collect a larger training data set that better reflects the possible variations that occur in naturalistic drinking episodes. This would allow us to train the model in a more robust way, and to better assess the model's generalization capabilities.

Appendix A List of Covariates

Name	Description
Session Type	Single, dual, or steady
Gender	Male or female
Ethnicity	African/African-American, Asian, caucasian, hispanics, or mixed
BMI	Body mass index
RM	Resting metabolism
SM	Skeletal muscle
Wais_Hip	waist-to-Hip ratio
AGE	Age of participants in years
ATE	Number of hours from eating to start of drinking session
WATER	Number of hours from time last drank water to time start drinking
CAL_LUN	Amount of calories for lunch
BPBase_S	Systolic blood pressure at rest
BPBase_D	Diastolic blood pressure at rest
BPPeak_S	Systolic blood pressure at peak
BPPeak_D	Diastolic blood pressure at peak
BPLow_S_S	Lowest systolic blood pressure reading before lunch
BPLow_S_D	Diastolic blood pressure reading at lowest systolic reading before lunch
BPLow_D_S	Systolic blood pressure reading at lowest diastolic reading before lunch
BPLow_D_D	Lowest diastolic blood pressure reading before lunch
28d_M	Past 28 days: maximum number of drinks on single day
28d_B	Number of binge days (5 drinks for women, 4 for men)
28d_T	Total number of drinks
90d_M	Past 90 days: maximum number of drinks on single day

Name	Description
90d_B	Number of binge days (5 drinks for women, 4 for men)
90d_T	Total number of drinks

Funding This research was supported in part by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) grant no. R01AA026368.

Data availability The datasets generated during and/or analyzed during the current study are not publicly available due to restrictions on human subject data, but are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethics approval This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by USC IRB.

References

1. Caddy G, Sobell M, Sobell L (1978) Alcohol breath tests: criterion times for avoiding contamination by "mouth alcohol''. Behav Res Methods Instrum 10:814–818
2. Lister RG, Gorenstein C, Risher-Flowers D, Weingartner HJ, Eckardt MJ (1991) Dissociation of the acute effects of alcohol on implicit and explicit memory processes. Neuropsychologia 29(12):1205–1212. https://doi.org/10.1016/0028-3932(91)90034-6
3. Sakai JT, Mikulich-Gilbertson S, Long RJ, Crowley TJ (2006) Validity of transdermal alcohol monitoring: fixed and self-regulated dosing. Alcohol Clin Exp Res 30:26–33. https://doi.org/10.1111/j.1530.0277.2006.00004.x
4. Gamella M, Campuzano S, Manso J, de Rivera GG, López-Colino F, Reviejo AJ, Pingarrón JM (2014) A novel non-invasive electrochemical biosensing device for in situ determination of the alcohol content in blood by monitoring ethanol in sweat. Anal Chimica Acta 806:1–7. https://doi.org/10.1016/j.aca.2013.09.020
5. Nyman E, Palmlöv A (1936) The elimination of ethyl alcohol in sweat. Acta Physiol 74(2):155–159. https://doi.org/10.1111/j.1748-1716.1936.tb01150.x
6. Pawan GLS (1968) Physical exercise and alcohol metabolism in man. Nature 218:966–967
7. Swift RM (2000) Transdermal alcohol measurement for estimation of blood alcohol concentration. Alcohol Clin Exp Res 24(4):422–423. https://doi.org/10.1111/j.1530-0277.2000.tb02006.x
8. Swift RM (1993) Transdermal measurement of alcohol consumption. Addiction 88(8):1037–1039. https://doi.org/10.1111/j.1360-0443.1993.tb02122.x
9. Swift RM, Martin CS, Swette L, LaConti A, Kackley N (1992) Studies on a wearable, electronic, transdermal alcohol sensor. Alcohol Clin Exp Res 16(4):721–725. https://doi.org/10.1111/j.1530-0277.1992.tb00668.x
10. Labianca DA (1990) The chemical basis of the breathalyzer: a critical analysis. J Chem Educ 67(3):259–261. https://doi.org/10.1021/ed067p259
11. Dai Z, Rosen IG, Wang C, Barnett N, Luczak SE (2016) Using drinking data and pharmacokinetic modeling to calibrate transport model and blind deconvolution based data analysis software for transdermal alcohol biosensors. Math Biosci Eng 13(5):911–934. https://doi.org/10.3934/mbe.2016023
12. Dougherty DM, Charles NE, Acheson A, John S, Furr RM, Hill-Kapturczak N (2012) Comparing the detection of transdermal and breath alcohol concentrations during periods of alcohol consumption ranging from moderate drinking to binge drinking. Exp Clin Psychopharmacol 20:373–381
13. Dougherty DM, Hill-Kapturczak N, Liang Y, Karns TE, Cates SE, Lake SL, Roache JD (2014) Use of continuous transdermal alcohol monitoring during a contingency management procedure to reduce excessive alcohol use. Drug Alcohol Depend 142:301–306
14. Dougherty DM, Karns TE, Mullen J, Liang Y, Lake SL, Roache JD, Hill-Kapturczak N (2015) Transdermal alcohol concentration data collected during a contingency management program to reduce at-risk drinking. Drug Alcohol Depend 148:77–84
15. Dumett MA, Rosen IG, Sabat J, Shaman A, Tempelman L, Wang C, Swift RM (2008) Deconvolving an estimate of breath measured blood alcohol concentration from biosensor collected transdermal ethanol data. Appl Math Comput 196(2):724–743. https://doi.org/10.1016/j.amc.2007.07.026
16. Rosen IG, Luczak SE, Hu WW, Hankin M (2013) Discrete-Time Blind Deconvolution for Distributed Parameter Systems with Dirichlet Boundary Input and Unbounded Output with Application to a Transdermal Alcohol Biosensor. In: 2013 Proceedings of the Conference on Control and its Applications (CT) pp. 160–167. https://doi.org/10.1137/1.9781611973273.22
17. Rosen IG, Luczak SE, Weiss J (2014) Blind deconvolution for distributed parameter systems with unbounded input and output and determining blood alcohol concentration from transdermal biosensor data. Appl Math Comput 231:357–376. https://doi.org/10.1016/j.amc.2013.12.099
18. Fairbairn CE, Kang D, Bosch N (2020) Using machine learning for real-time bac estimation from a new-generation transdermal biosensor in the laboratory. Drug Alcohol Depend. https://doi.org/10.1016/j.drugalcdep.2020.108205
19. Oszkinat C, Shao T, Wang C, Rosen IG, Rosen AD, Saldich EB, Luczak SE (2022) Obtaining blood and breath alcohol concentration from biosensor measured transdermal alcohol level: Estimation and uncertainty quantification based on a covariate-dependent hidden markov model. Inverse Probl 38(5):005002
20. Oszkinat C, Luczak SE, Rosen IG (2021) Uncertainty quantification in the estimation of blood alcohol concentration using physics-informed neural networks. IEEE Trans Neural Netw Learn Syst
21. Leffingwell TR, Cooney NJ, Murphy JG, Luczak S, Rosen G, Dougherty DM, Barnett NP (2013) Continuous objective monitoring of alcohol use: twenty-first century measurement using transdermal sensors. Alcohol Clin Exp Res 37(1):16–22
22. Luczak SE, Ramchandani VA (2019) Special issue on alcohol biosensors: development, use, and state of the field: summary, conclusions, and future directions. Alcohol 81:161–165
23. Fairbairn CE, Kang D (2019) Temporal dynamics of transdermal alcohol concentration measured via new-generation wrist-worn biosensor. Alcohol Clin Exp Res 43(10):2060–2069
24. Hawekotte K, Luczak SE, Rosen IG (2021) Deconvolving breath alcohol concentration from biosensor measured transdermal

alcohol level under uncertainty: a bayesian approach. Math Biosci Eng 18(5):6739–6770
25. Yao M, Luczak SE, Rosen IG (2021) Tracking and Deconvolving Blood Alcohol Concentration from Transdermal Alcohol Biosensor Measurements: A Population Model-Based LQG Approach in Hilbert Space
26. Banks HT, Ito K (1988) A unified framework for approximation in inverse problems for distributed parameter systems. Control Theor Adv Technol 4(1):73–90
27. Tanabe H (1979) Equations of Evolution. In: Monographs and Studies in Mathematics, Pitman https://books.google.com/books?id=Dn6zAAAAIAAJ
28. Pazy A (1983) Semigroups of Linear Operators and Applications to Partial Differential Equations. In: Applied Mathematical Sciences, Springer, https://books.google.co.in/books?id=80XYPwAACAAJ
29. Banks HT, Kunisch K (2012) Estimation Techniques for Distributed Parameter Systems. Springer, https://www.springer.com/us/book/9780817634339
30. Yao M, Luczak SE, Rosen IG (2021) A population model-based lqg compensator for the control of intravenously-infused alcohol studies with transdermal sensing
31. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural comput 9:1735–80. https://doi.org/10.1162/neco.1997.9.8.1735
32. Hochreiter S (1991) Untersuchungen zu dynamischen neuronalen netzen
33. Raissi M, Perdikaris P, Karniadakis GE (2017) Physics informed deep learning (Part I): Data-driven Solutions of Nonlinear Partial Differential Equations
34. Raissi M, Perdikaris P, Karniadakis GE (2017) Physics informed deep learning (Part II): Data-driven Discovery of Nonlinear Partial Differential Equations
35. Raissi M, Perdikaris P, Karniadakis GE (2018) Multistep Neural Networks for Data-driven Discovery of Nonlinear Dynamical Systems
36. Raissi M, Perdikaris P, Karniadakis G (2018) Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J Comput Phys. https://doi.org/10.1016/j.jcp.2018.10.045
37. Sirlanci M, Rosen IG, Luczak SE, Fairbairn CE, Bresin K, Kang D (2018) Deconvolving the input to random abstract parabolic systems; a population model-based approach to estimating blood/breath alcohol concentration from transdermal alcohol biosensor data. Inverse problems **34**(12) . arXiv:1807.05088v1
38. Sirlanci M, Luczak SE, Fairbairn CE, Kang D, Pan R, Yu X, Rosen IG (2018) Estimating the distribution of random parameters in a diffusion equation forward model for a transdermal alcohol biosensor. Automatica. to appear, arXiv:1808.04058
39. MacKay DJC (2003) Information Theory Inference and Learning Algorithms. Cambridge University Press, Cambridge
40. Gal Y, Ghahramani Z (2015) Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. arXiv:1506.02142
41. Gal Y (2016) Uncertainty in deep learning. PhD thesis, University of Cambridge, Cambridge
42. Caflisch R (1998) Monte carlo and quasi-monte carlo methods. Acta Numerica 7:1–49. https://doi.org/10.1017/S0962492900002804
43. Hinton GE, van Camp D (1993) Keeping the neural networks simple by minimizing the description length of the weights. In: Proceedings of the Sixth Annual Conference on Computational Learning Theory. COLT '93, pp. 5–13. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/168304.168306. https://doi.org/10.1145/168304.168306
44. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(56):1929–1958
45. Saldich EB, Wang C, Rosen IG, Goldstein L, Bartroff J, Swift RM, Luczak SE (2020) Obtaining high-resolution multi-biosensor data for modeling transdermal alcohol concentration data. Alcoholism: Clinical and Experimental Research **44**
46. Banks HT, Burns JA, Cliff EM (1981) Parameter estimation and identification for systems with delays. SIAM J Control Optim 19(6):791–828. https://doi.org/10.1137/0319051
47. Casella G, Berger RL (2002) Statistical Inference vol. 2. Duxbury, Pacific Grove, CA https://books.google.com/books/about/Statistical_Inference.html?id=0x_vAAAAMAAJ
48. Kingma D, Ba J (2014) Adam: A method for stochastic optimization. In: International Conference on Learning Representations
49. The Shapley Value (1988) Essays in Honor of Lloyd S. Shapley Cambridge University Press. https://doi.org/10.1017/CBO9780511528446