Physiologically-Informed Gaussian Processes for Interpretable Modelling of Psycho-Physiological States

Shadi Ghiasi[1]*, Andrea Patane[2]*, Luca Laurenti[2], Claudio Gentili[3], Enzo Pasquale Scilingo[1], *Member, IEEE*, Alberto Greco[1], *Member, IEEE* and Marta Kwiatkowska[2]

***Abstract*— The widespread popularity of Machine Learning (ML) models in healthcare solutions has increased the demand for their interpretability and accountability. In this paper, we propose the *Physiologically-Informed Gaussian Process* (*PhGP*) classification model, an interpretable machine learning model founded on the Bayesian nature of Gaussian Processes (GPs). Specifically, we inject problem-specific domain knowledge of inherent physiological mechanisms underlying the psycho-physiological states as a prior distribution over the GP latent space. Thus, to estimate the hyper-parameters in *PhGP*, we rely on the information from raw physiological signals as well as the designed prior function encoding the physiologically-inspired modelling assumptions. Alongside this new model, we present novel interpretability metrics that highlight the most informative input regions that contribute to the GP prediction. We evaluate the ability of *PhGP* to provide an accurate and interpretable classification on three different datasets, including electrodermal activity (EDA) signals collected during emotional, painful, and stressful tasks. Our results demonstrate that, for all three tasks, recognition performance is improved by using the *PhGP* model compared to competitive methods. Moreover, *PhGP* is able to provide physiological sound interpretations over its predictions.**

***Index Terms*— Gaussian Process, Interpretable Machine Learning, Electrodermal activity, Bayesian learning, Psycho-physiology**

I. Introduction

Owing to their success in several application domains [1], Machine Learning (ML) models are becoming the method of choice for tackling recognition and detection problems in healthcare applications such as affective computing [2]. However, to validate ML models in healthcare domains, traditional ML metrics for performance assessment, e.g., accuracy, precision or recall, are no longer sufficient. Alongside these metrics, the interpretability of model predictions is the highest priority for clinicians and healthcare practitioners, as it allows them to understand, debug and assess the predictive ability of ML model [3]. For instance, in real-life affective computing applications, a psychologist may want to know why a ML model is diagnosing a mental condition, or a physiologist may be interested in understanding whether the physiological dynamics is taken into account in the learning algorithm [4]. Therefore, the choice of the best model for recognition of psycho-physiological states is subject to a trade-off between the performance and the interpretability of the ML system.

By working with the raw, unprocessed signal, *end-to-end* models, in particular, have often been found to outperform human/expert recognition performance. Unfortunately, the resulting models behave as black-boxes and are less interpretable at the model design level. Furthermore, because of the reliance on human and expert participation in data collection experiments, affective datasets are often small in size, sparsely labelled, and thus violate the big-data assumption that deep learning relies on [5].

On the other hand, rather than employing end-to-end models, practitioners often rely on standard and well understood feature extraction techniques [6] that utilise human-generated expert knowledge and established problem-specific findings. While previous literature has considered learning end-to-end models in conjunction with feature-extraction pipelines [7], [8], to the best of our knowledge an effective method for their principled and interpretable integration in the context of affective computing is missing. Intuitively, incorporating expert domain-specific knowledge into an end-to-end framework has the potential of further informing a ML model about the context of the input data. This may not only enhance the recognition performance, thus enabling end-to-end learning, but can also provide the practitioner with a more transparent, understandable ML model.

In this work, we introduce the *Physiologically-Informed Gaussian Process* (*PhGP*) model as a tool for integrating, through a Bayesian principled approach, information contained in automatically-discovered patterns in the raw signal data with expert domain knowledge available about the problem at hand. Specifically, we proceed by encoding the latter, assumed to be in the form of probabilistic assumptions about the data-generating physiological process, as a set of stochastic objects that are probabilistically correlated with the input raw physiological signal and the subject's psycho-physiological state. By relying on a MAP (Maximum *A-Posteriori*) estimation for the quantities involved, we show how this can be used to infer a prior distribution over a Gaussian Process (GP) classification model.

Intuitively, in the *PhGP* framework, the prior

[1] Department of Information Engineering & Bioengineering and Robotics Research Centre E. Piaggio, School of Engineering, University of Pisa, Italy.

[2] Department of Computer Science, University of Oxford.

[3] Department of General Psychology, University of Padua, Italy.

This project was partially funded from the EU's Horizon 2020 program under the Marie Sklodowska-Curie grant agreement No. 722022 "AffecTech", from the Italian Ministry of Education and Research (MIUR) in the framework of the CrossLab project (Departments of Excellence), and from the ERC under the European Union's Horizon 2020 research and innovation programme FUN2MODEL (grant agreement No. 834115).

* These authors contributed equally to this work.

Corresponding author email: shadi1337@gmail.com

physiologically-based model is thus used as a soft starting point for model training, and is adapted in a probabilistic fashion according to the dataset. This potentially reduces the risk of over-fitting when learning the model directly from the raw signal and in practice allows for learning even with affective datasets of small size.

An additional aspect that enhances the interpretability of recognition models from affective biomarkers is ranking and selecting those salient features that make a major contribution to the classifier's decision. This may be potentially of interest to a clinician or physiologist since it allows better understanding of the physiological patterns underlying the classification problem [9]. One example is a recently proposed interpretation of deep regression models for depression detection by identifying salient regions in face images in terms of their severity level, which reveal the visual depression patterns on faces [4].

We take a similar approach and, to enhance the interpretability of our *PhGP* framework, in addition to the transparency, formulate a novel methodology to produce quantitative and visual explanation of *PhGP* predictions by generating physiological activation maps (PAMs), which represent the salient patterns leading to the classifier's decisions.

We implement and empirically validate our methods on three datasets of psycho-physiological state recognition from the Electrodermal Activity (EDA) signal, relying on the Convex optimisation tool for EDA processing (cvxEDA) [10] for the design of the physiological prior model. In particular, we focus on the DEAP (A Dataset for Emotion Analysis using EEG, Physiological and Video Signals) dataset for recognition of video-induced emotion [11], the BHVP (BioVid Heat Pain) dataset for pain recognition [12] and Stroop (a dataset containing EDA signals collected during a paced stroop task test for stress recognition).

Empirical results demonstrate that, by combining the raw signal with the physiologically-based prior function, *PhGP* outperforms GP-based models that have access to only one of these sources of information. Further comparison of *PhGP* with a state-of-the-art classification method, the support vector machines embedded with recursive feature elimination (*SVM-RFE*) demonstrates that the former obtains competitive performance results, while still providing physiologically sound interpretations over its predictions.

This paper is a significant extension of our previous work [13], which we have extended in a number of directions.

- We develop techniques that, by building on physiologically-inspired priors explored in [13], enable us to train GP classification models directly from raw, unprocessed physiological signals while providing a more transparent and explainable GP architecture.
- We develop methods to provide quantitative and visual interpretation of *PhGP* predictions by relying on the explicit formulation of the GP inference equations, in addition to the explainability of our framework at the model design level.
- We provide empirical comparison of *PhGP* with standard GP models and a state-of-the-art feature-based classification algorithm on three datasets of psycho-physiological state recognition.

II. Related Work

Affective recognition from physiological sensors, i.e., the problem of inferring user's emotional/affective state from signals recorded from one's body, is routinely achieved through extraction and processing of features [6]. Mathematical models have been specifically developed as a means to uncover and make explicit the relationship that exists between the psycho-physiological state of a user and his/her body signals. Examples include the *integral pulse frequency modulation* [14] and the *point-process* model [15] for the modelling of heart rate variability dynamics; *causal modelling* [16] and *cvxEDA* [10] for explaining EDA dynamics; as well as the *recursive penalised least squares* approach for the solution of the electroencephalogram (EEG) signal [17]. Compared to generic feature extraction methods, model-based techniques mathematically encode domain-specific expert knowledge about the physiology of the affective modelling problem itself, and as such are able to provide detailed explanations of the inherent physiological mechanisms [16].

End-to-end learning, especially in the form of deep neural network models, has consistently outperformed standard ML pipelines for affective computing, at least in cases where a sufficient amount of labelled data is available [8], [18]–[21]. Unfortunately, overfitting problems and the lack of interpretability inherent in neural networks has thus far limited the use of these methods in practical clinical applications [5]. In an effort to overcome these issues, several works have considered techniques for extensive data augmentation [21], [22] and transfer learning or pre-trained networks [19], [23], as well as learning based on hand-crafted features [24] or creating ensembles of deep and shallow models [8]. While mitigating these issues, data augmentation and transfer learning do not fundamentally overcome them, and the use of hand-crafted features restricts *a-priori* the learning capabilities of deep models. On the other hand, our method, by relying on patterns automatically learned from raw data by the GP, reduces the risk of overfitting by probabilistically centering the model around the explicit solution given by a physiologically-inspired approach.

GP models have been applied in various forms in physiological signal analysis [25]–[28]. Beside GP models, the Relevance Vector Machines are another type of probabilistic extended linear models which offer a higher flexibility for the choice of basis functions with prior on weights that enforces sparse solutions [29], [30]. However, physiologically-based design of the prior distribution in the Bayesian architecture of GP models has not been fully investigated, and priors used in the literature tend to be uninformative. The authors in [25] proposed an approach for designing priors for GPs specifically tailored to capturing hemodynamics in functional magnetic resonance imaging analysis, showing that an informed GP model significantly outperforms a GP trained on uninformative priors. Similarly, the authors in [27] proposed a pseudo-Bayesian method for the estimation of intracranial pressure, where the model likelihood is informed and adapted by physiological modelling. We build on this literature to design an approach, in which the posterior distribution is informed both by the peculiarities of the dataset at hand and the information embedded within mathematical physiological models.

In this work we focus on applying our techniques to EDA signals, but note that the approach can also be used for other physiological signals. Several studies have suggested that the EDA signal, even in single-modality settings, can provide objective means of assessing psycho-physiological states, including emotional changes and distress associated with pain [6], [19], [31], [32].

III. Methods

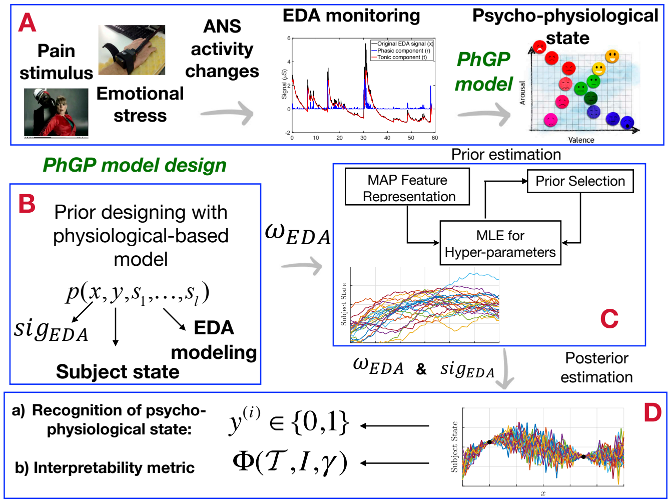

Fig. 1: *PhGP* recognition pipeline implemented for the classification of psycho-physiological states. A) The ANS is monitored by acquiring the EDA signal. B) The prior function is estimated based on the EDA physiological-based model. C) Within GP the posterior is estimated. D) EDA signals are classified along with interpretability metrics.

A depiction of the complete prediction pipeline for the *PhGP* model is given in Figure 1, in the case of recognition of psycho-physiological states from EDA recordings (block A of the plot). First, information from a physiologically-based model of the EDA signals is encoded into a probabilistic generative model that captures their relationship with the raw input signal and the subject's psycho-physiological state (block B in the plot, and discussed in Section III-B). This is then used, together with a MAP estimation of the feature representation and an approximate Maximum Likelihood Estimation (MLE) for the hyper-parameters, to define a Gaussian prior over the *PhGP* model (block C in the plot). Finally, posterior Bayesian inference is performed for the *PhGP* model to obtain a prediction for the subject's psycho-physiological state $y^{(i)} \in \{0, 1\}$, and interpretability analysis is employed to provide a quantitative explanation of the prediction made (block D in the plot, and discussed in Section III-C).

A. Preliminaries

We consider a generic training dataset associated to a binary classification problem, $\mathcal{D} = \{(x^{(i)}, y^{(i)}) \mid x^{(i)} = [x_1^{(i)}, \ldots, x_n^{(i)}] \in \mathbb{R}^n, y^{(i)} \in \{0, 1\},\ i = 1, \ldots, N\}$, where $N = |\mathcal{D}|$ represents the number of observation data points. Each $x^{(i)}$ denotes the $i-$th, raw, physiological signal recorded at n discrete time-steps, while $y^{(i)}$ represents its associated class label, i.e., the subject's psycho-physiological state that we aim to model. Let $\mathbf{x} = [x^{(1)}, \ldots, x^{(N)}] \in \mathbb{R}^{N \times n}$ be the combined vector of input physiological signals and $\mathbf{y} = [y^{(1)}, \ldots, y^{(N)}] \in \mathbb{R}^N$ be its associated class vector, encoding the psycho-physiological state of interest (see Section IV-A).

In two-class classification we proceed by defining a latent variable, $f \in \mathbb{R}$, that represents the classification *logit*, and relate it to the class probability by means of a sigmoid function $\sigma(\cdot)$.[1] We employ Gaussian Process classification with Laplace approximation for modelling the relationship between $\mathbf{x}$ and $\mathbf{y}$. Briefly, this is achieved by placing a Gaussian prior function, $p(f|x)$, over the latent variable, performing Bayesian inference on it, and finally computing the predictive probability function, denoted with $\pi(y = 1|\mathcal{D}, x)$, that encodes the probability that x belongs to class 1.

The key observation that we exploit in this paper is that the prior function $p(f|x)$ depends on a mean function $\mu : \mathbb{R}^n \longrightarrow \mathbb{R}$ and a kernel (covariance) function $k : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$. In Section III-B we show how these can be adapted to automatically fit data arising from psycho-physiological processes. An in-depth overview of Gaussian Process classification can be found in [33], and a summary of the relevant background is given in the Supplementary Material.

B. The Physiologically-Informed Gaussian Process Model

The *PhGP* model builds on Bayesian learning in order to embed information from a physiological model into the training process. To achieve this, we encode the model as a functional distribution over the latent variable $f \in \mathbb{R}$ and feed it into the definition of the GP prior, $p(f|x)$.

Specifically, in *PhGP* we interpret a physiological model as a set of unobservable sub-processes $\mathbf{s} = [s_1, \ldots, s_l]$. Given a subject's psycho-physiological state, y, the process $\mathbf{s}$ gives rise to the observable physiological signal x according to a stochastic generative model of the form:

$$[x, y, \mathbf{s}] \sim p(x, y, s_1, \ldots, s_l) \qquad (1)$$

for an unknown density function p. Intuitively, $\mathbf{s}$ captures the physiological phenomenon related to certain physiological processes. Most of these, are not actually directly observable through the use of physiological signal monitoring, hence we assume that $\mathbf{s}$ is unobservable. For example, in the case of the EDA signal (discussed in more detail in Section III-D), a physiological model, $p(x, y, s_1, \ldots, s_l)$, captures the relationship between the subject's condition and the sudomotor nerve activity (SMNA) that gives rise to a variation of the skin conductance properties. The idea is that the space in which the sub-processes contained in $\mathbf{s}$ are defined allows for a better understanding of the signal properties, and for the extraction of a set of m relevant quantifiers (features), which we denote as

[1] This can be extended to multi-class classification by using either a one-vs-all classification approach or multi-output likelihood functions.

$\omega(\mathbf{s}) = [\omega_1(\mathbf{s}), \ldots, \omega_m(\mathbf{s})] \in \mathbb{R}^m$. In this way, the relationship with the subject's condition, y, is understood in terms of direct, physiological correlation.

In *PhGP*, we consider the feature vector $\omega(\mathbf{s})$ as a building block of the prior knowledge used to approximate the effect of $\mathbf{s}$ on the generation of the subject's condition y in Equation (1). We do this by embedding this information in the prior mean function, $\mu(x)$, of the GP and employ a parametric approach to estimate its effect on the GP output. In particular, we investigate the suitability of polynomial and trigonometric functions of the form:

$$m_1(\mathbf{s}|\alpha) = \sum_{p=0}^{d} \sum_{j=1}^{m} \alpha_{pj} \omega_j(\mathbf{s})^p \tag{2}$$

$$m_2(\mathbf{s}|\alpha) = \sum_{p=1}^{d} \alpha_p^{(1)} \cos\left(\sum_{j=1}^{m} \alpha_{pj}^{(2)} \omega_j(\mathbf{s}) + \alpha_p^{(3)} \right), \tag{3}$$

where α is a vector of unknown parameters that adapt the shape of $m_j(\mathbf{s}|\alpha)$.

Parameter d in Equation (2) is the degree of the polynomial function and we retrieve a constant, linear, quadratic and cubic function for the values $d = 0, 1, 2, 3$. In Equation (3) d is the number of projected cosine components.

We observe that, by relying on the probabilistic relationship that exists between x and $\mathbf{s}$ (Equation (1)), $m_j(\mathbf{s}|\alpha)$ can be used to naturally induce a prior mean function over the GP. By marginalising over the random variable $\mathbf{s}$ we in fact obtain:

$$\mu(x|\alpha) = \int m_j(\mathbf{s}|\alpha) p(\mathbf{s}|x) d\mathbf{s}. \tag{4}$$

Notice that the resulting mean prior function $\mu(x|\alpha)$ we obtain, combines standard parametric mean functions (e.g., those of Equations (2) and (3)) with the information coming from the problem specific physiological model of Equation (1) in a modular fashion. In general, this integral cannot be computed analytically, so in practice we employ a Monte Carlo approximation for its computation, namely:

$$\mu(x|\alpha) \approx \sum_{i=1}^{M} m_j(\mathbf{s}_i|\alpha) p(\mathbf{s}_i|x) \tag{5}$$

for M random samples of $\mathbf{s}$. The *PhGP* prior is then defined by the choice of the kernel function $k(x, x)$, for which we employ the squared-exponential kernel computed directly on the raw physiological signal x, as this provides flexible and smooth GP models that can adapt to different classification boundaries for specific choices of hyper-parameters [34].

The mean and covariance thus defined centre the *PhGP* prior around the model estimation provided by the physiologically-based generative model. The learning procedure for the *PhGP* model then follows the lines outlined in the Supplementary Material for the standard GP case.

In particular, by plugging Equation (5) for the physiologically informed prior mean for the a-posteriori GP mean, for a test point x^*, in *PhGP* we obtain:

$$\hat{\mu}(x^*) = \sum_{i=1}^{M} m_j(\mathbf{s}_i|\alpha) p(\mathbf{s}_i|x^*) + \mathbf{k}^{*T} K^{-1} \hat{\mathbf{f}} \tag{6}$$

where $\mathbf{k}^* \in \mathbb{R}^N$ is the covariance computed between x^* and each point in the training set, $K \in \mathbb{R}^{N \times N}$ is the training set covariance matrix, and $\hat{\mathbf{f}} \in \mathbb{R}^N$ is the mode of the posterior distribution over the training set. Further notice that $p(\mathbf{s}_i|x^*)$ represents the conditional density function of $\mathbf{s}$ given x^*, evaluated at the Monte Carlo sample $\mathbf{s}_i$, and $m_j(\mathbf{s}_i|\alpha)$ is the evaluation of the function m_j with parameters α in $\mathbf{s}_i$. We note that, in Equation (6), the solution provided by the physiological model is adapted by the raw data naturally following the Bayes rule.

C. *Interpretability Analysis of the* PhGP *Model*

A key advantage of relying on physiologically-informed features in *PhGP* is the ease of interpretability of the resulting model architecture. This is because the features are directly fed into the learning process of the GP classification model. However, an additional degree of interpretability can be achieved after obtaining the predictions made by the model. Therefore, we build an interpretability framework for our proposed $PhGP$ model. We achieve this by extending the interpretability methods discussed in [35], [36] based on the explicit form of the GP inference equations to the context of *PhGP* modelling. Namely, we proceed by propagating small input perturbations in succession through the physiologically-based prior model and then through the GP posterior in order to obtain an estimate of the contribution of each data point to the overall model prediction. The outcome is an interpretability metric, denoted Φ, that corresponds to the importance of each data point in the recognition task.

Consider a generic input point, $x \in \mathbb{R}^n$, then, for any subset of indices $I \subseteq \{1, 2, \ldots n\}$, we call x_I the subvector of x that includes only the indices of I, that is, $x_I = [x_i]_{i \in I}$. For an input point x^*, a subset of indices I, a norm $|\cdot|$ and a radius $\gamma > 0$, we define $T^I_{\gamma,x^*} = \{x \in \mathbb{R}^n \text{ s.t. } |x_I - x^*_I| \leq \gamma\}$ as the set of allowed perturbations of magnitude up to γ around x^*. In the following definition we quantify the maximum effect that local perturbations of the subvector of indices I have on the classification probabilities.

Definition 1: Consider T^I_{γ,x^*}, then we define

$$\phi(T^I_{\gamma,x^*}) = \max_{x \in T^I_{\gamma,x^*}} \pi(y = 1|\mathcal{D}, x) - \min_{x \in T^I_{\gamma,x^*}} \pi(y = 1|\mathcal{D}, x).$$

Then, for a finite set of test points $\mathcal{T}$ we define the interpretability metric by $\Phi(\mathcal{T}, I, \gamma) = \frac{1}{|\mathcal{T}|} \sum_{x^* \in \mathcal{T}} \phi(T^I_{\gamma,x^*})$.

Intuitively, for a test point x^*, $\phi(T^I_{\gamma,x^*})$ is a measure of how much local perturbations of the indices I of x^* can change the classification probabilities. $\Phi(\mathcal{T}, I, \gamma)$ is the average of ϕ over a set of input points, that is, $\Phi(\mathcal{T}, I, \gamma)$ measures how much on average the perturbations of a test point will affect the classification probabilities. Details of the computation of the interpretability metric Φ for *PhGP* models are given in the Supplementary Material.

D. *EDA-based Modelling*

EDA broadly refers to the variations in the skin conductance induced by the sudomotor nerve activity which modulates the sweat secretion of the eccrine glands. It can be measured through an EDA meter, a device that displays the electrical conductance change between two points over time. We can now give an explicit formulation for EDA-based *PhGP* modelling, by considering the *cvxEDA* model [10] as the generative model of the physiological signal. Namely, this represents the observed n-sample long signal x as a sum of three n-long components, a tonic component ($s_1 = t$), a phasic component ($s_2 = r$) and an additive noise term (ϵ), according to the following equation:

$$x = r + t + \epsilon. \tag{7}$$

The tonic activity contains information about the overall psycho-physiological state of the subject, while the phasic component shows rapid changes in EDA signals directly related to an external physiological stimulation. The phasic component is the output of the convolution between the SMNA and an impulse response function that describes the sweat diffusion process. We refer to the sparse SMNA driver of the phasic component as p. We encode the parameters of this model in the stochastic generative model in the form of Equation (1). We then use the following standard quantifiers of x, t, r and p to form a set of features that constitute the vector $\omega_{EDA}(\mathbf{s})$ [10], [37]:

- $\omega_{r,p}(\mathbf{s})$: includes the number of significant phasic driver peaks ($nSCR$), the sum of Skin Conductance Response (SCR) amplitudes ($SumAmpSCR$), the maximum value of SCR amplitudes ($MaxAmpSCR$), and the mean and standard deviation of phasic activity ($PhasicMean$, $PhasicStd$);
- $\omega_t(\mathbf{s})$: includes mean and standard deviation of tonic activity ($TonicMean$,$TonicStd$);
- $\omega_x(\mathbf{s})$: includes $EDASymp$, which is highly correlated to the activity of the sympathetic nervous system and is obtained by integrating the spectrum of x within the $(0.045 - 0.25Hz)$ frequency band.

IV. Experimental applications

A. Experimental Data

We perform our analyses on two publicly available datasets, which report EDA signals during changes in autonomic nervous system (ANS) activity, namely, the DEAP dataset [11], the BHVP (BioVid Heat Pain) dataset [12] and the Stroop test, a dataset collected in our laboratory (block A in Figure 1). Details of all three datasets are given in the following paragraphs.

1) Affective valence recognition (emotional stimulus): The DEAP dataset consists of multi-modal physiological recordings (including EDA), recorded from 32 healthy subjects watching different affective video clips. During each trial, the index of the current trial was first shown for 2 seconds and a consecutive 5 seconds of baseline recording was followed. Then, the subjects were exposed to the emotional stimulus for 1 minute. Finally, they were asked to mark the stimulus on a scale of 1-9. Since the 32 initial subjects were recorded by means of two different EDA acquisition systems, as reported in the dataset description page, we select only the first 21 subjects, i.e., the largest group recorded with the same system to avoid a bias that was evident from the preliminary analysis of the signals. In this paper we focus on the highest arousal and highest valence videos and choose the data recorded during the 5 highest positive valence/highest arousal and 5 highest negative valence/highest arousal videos of the subjects. The resulting dataset contains 105 observations equally balanced between the two class. Additional details of this dataset can be found in [11].

2) Autonomic arousal recognition (pain stimulus): In the BVHP dataset, a group of 87 subjects underwent a heat-induced pain experiment of four different intensities, while their physiological response was being recorded (including EDA signal). Each pain stimulus was applied at the subject's right arm for around 5 seconds. Each of the specific pain level stimulus was elicited 20 times in a randomised order for each study participant. There was a randomised rest of 8 to 12 seconds between the stimuli. We choose two states corresponding to the states with the highest and the lowest level of heat pain stimulus representing two diverse psycho-physiological states in subjects. This choice was according to pre-existing research on the same dataset, which enables baseline comparison of our results with the literature [38], [39]. We then build the training set related to this dataset from 174 observations in each class. Additional details of this dataset can be found in [12].

3) Stroop test: 33 healthy subjects volunteered to take part in this study in University of Pisa. The experimental protocol consisted of a 5 minute resting state followed by a stressor, namely, the paced stroop test lasting for 3 minutes [40]. During this task, the subjects were shown words whose meaning was different from their displayed colors. The subjects had two seconds to press the button corresponding to the color of the displayed word and not the corresponding meaning. In case of any mistakes or missed answer, a buzzer was activated and the counter showing the number of consecutive correct answers would turn back to zero. During the experiment, the EDA signal was monitored. The subjects gave their written informed consent and the experiment was approved by the "Comitato Etico Regionale per la Sperimentazione Clinica della Regione Toscana", section "Area Vasta Nord Ovest" - Protocol n. 7803, Registry number 1072, approved on 18 Jan 2018. The recordings were carried out in agreement with the Declaration of Helsinki.

B. Recognition Pipeline with the PhGP Model

A key advantage of *PhGP* is that, in view of its probabilistic formulation, the model predictions take into account the information provided by the physiologically-based model as well as the raw signals available in the given dataset. In the experiments discussed in Section V, we compare *PhGP* with variants in which only one of the two sources of information is available, and specifically the following:

1) The input data of the classification model are raw physiological signals x. We refer to this as *Raw-GP*.
2) The classification model is the last step of a feature extraction pipeline associated to the feature vector $\omega \in \mathbb{R}^m$. We refer to this as *Feat-GP*.

The *PhGP* model can be viewed as a combination of the two approaches, as it incorporates both learning with GPs and the knowledge of the features of the physiological model. The training of the *PhGP* model proceeds from raw data by adapting the model distribution around the explicit solution of the physiologically-inspired computational model. We investigate the parametric prior functions described in Section III-B for all three GP-based models. Moreover, for comparison with well performing classification methods outside the GP context, we also evaluate the *SVM-RFE* algorithm [41] on the experimental data (see the Supplementary Material for additional details on this method). MATLAB software (R2017b version) and the Gaussian Processes for Machine Learning (GPML) toolbox [34] were used to implement GP model training and prior function estimation.

C. Interpretability pipeline of PhGP model

After training the *PhGP* model as well as the other two variants (*Raw-GP* and *Feat-GP*), the interpretability metric (Φ) is estimated for each input data point of each recognition model (refer to Section III-C). This metric indicates:

1) For *Raw-GP*: the contribution of each data sample in the input signal in the final prediction.
2) For *Feat-GP*: the contribution of each feature in the feature vector ω in the final prediction.
3) For *PhGP*: the contribution of each data sample in the input signal in the final prediction. However, inherently, when performing posterior inference for *PhGP*, the contribution of the physiologically-based features is confounded with those of the raw data from the input dataset.

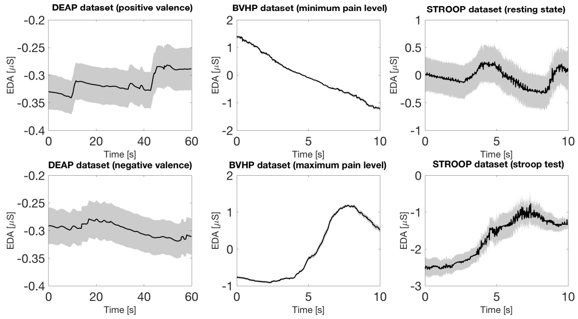

Fig. 2: Group-wise EDA dynamics along the timeline for each psycho-physiological condition in each dataset. Note that the continuous black line represents the $Median$ value and the gray area represents the MAD (refer to Section V for the definition of MAD) along all subjects.

V. Results

Figure 2 shows the EDA trends averaged across all subjects in the two psycho-physiological conditions for each of the three applications. The plots in this figure are expressed as $Median \pm 1.4826MAD(X)/n$, (where $MAD(X) = Median(|X - Median(X)|)$, with X as the EDA signal and n as the number of subjects in each dataset) over time.

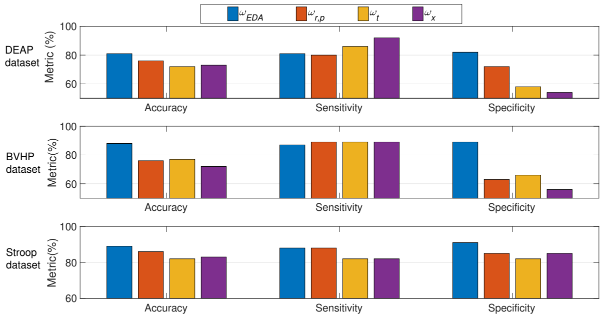

Fig. 3: Comparative performance (sensitivity, specificity and accuracy) of *PhGP* with different subsets of features in the prior function for each dataset. Refer to Section III-D for the definition of each subset.

A. Recognition Results

In Table I we provide recognition results of the *PhGP* model and compare its performance with that of *Raw-GP* and *Feat-GP* as well as the *SVM-RFE* algorithm. In particular, we investigate the performance of the GP-based models with respect to different choices of the parametric form of the mean prior functions (i.e., zero, constant, linear, quadratic, cubic and trigonometric). We do not give results for *PhGP* with zero and constant mean, as the *PhGP* formulation relies on a non-trivial mean function. The results reported are computed through a Leave-One-Subject-Out (LOSO) cross-validation procedure, so that the results and the models obtained are subject-independent. Namely, at each iteration of the LOSO validation scheme, the recognition model is trained using data from $M-1$ subjects (where M is the total number of subjects) and tested on the data from the left-out subject. This procedure is iterated M times. In the table we report final performance results averaged over all subjects in terms of sensitivity (i.e., number of true positive assessments/ number of all positive assessments) , specificity (i.e., number of true negative assessments/ number of all negative assessments) and accuracy (i.e., number of correct assessments/ number of all assessments) of predictions.

The results reported in Table I suggest that the *PhGP* model obtains an overall higher accuracy for all GP prior functions compared to the *Feat-GP* and the *SVM-RFE* model in all three datasets, while it outperforms the *Raw-GP* model for some specific GP prior functions. For example, with the linear prior function, *PhGP* obtains 3%, 2% and 12% higher accuracy compared to *Feat-GP* and 13%, 2% and 1% higher accuracy compared to the *Raw-GP* model in DEAP, BVHP and Stroop datasets, respectively.

Results obtained from training the *PhGP* model with the linear, quadratic, cubic and trigonometric prior functions show virtually similar performance in the BVHP dataset (1% difference in accuracy). However, the difference in performance is more evident in the DEAP dataset ($6-9\%$) and the Stroop dataset ($3-7\%$). Interestingly, the highest accuracy for the *PhGP* model is achieved with the linear function in all datasets.

Observe that *Feat-GP* significantly outperforms *Raw-GP*

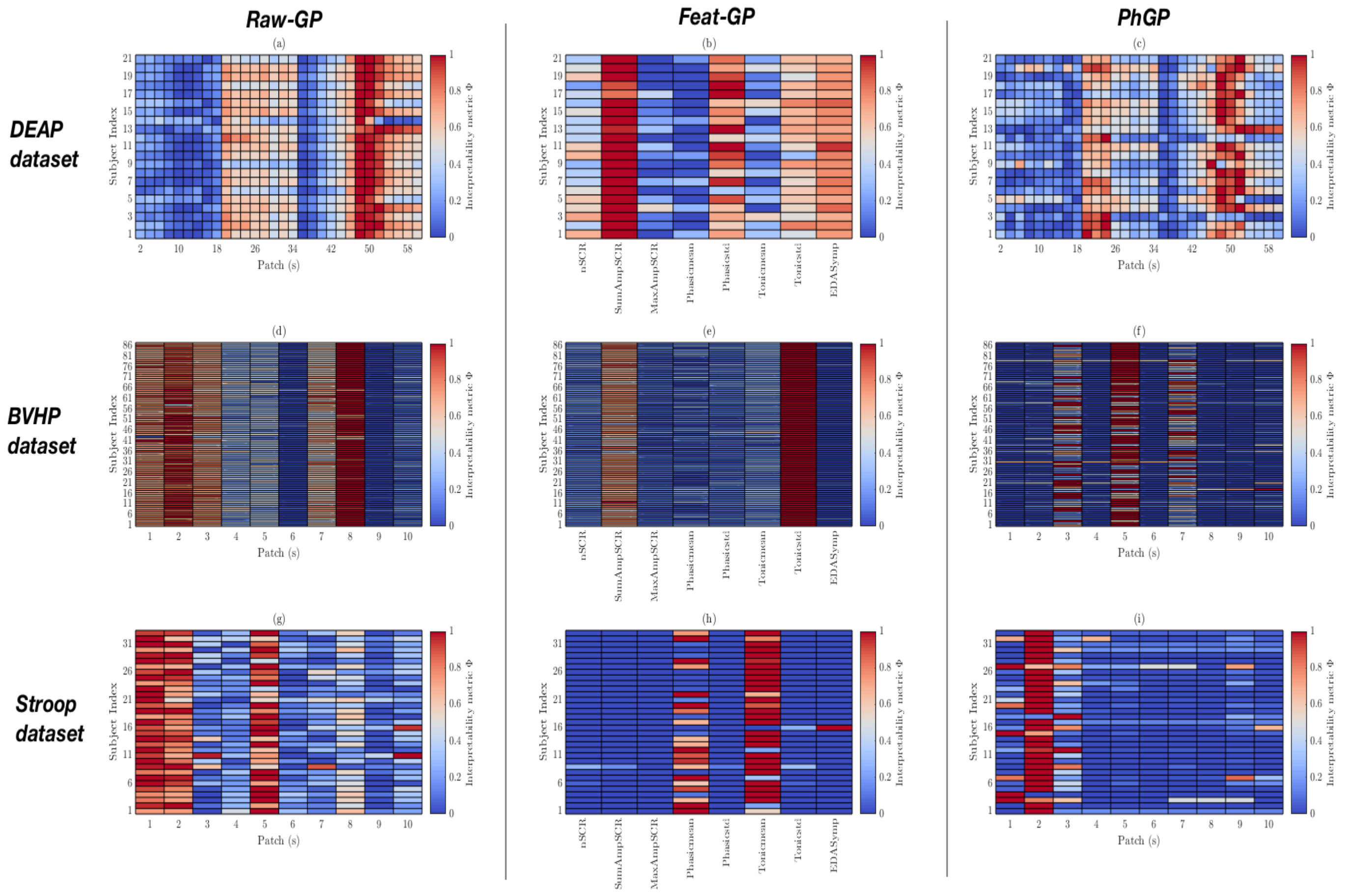

Fig. 4: (a),(d),(g): Physiological activation maps (PAM) displaying the contribution of each data patch in DEAP, BVHP, Stroop datasets for the *Raw-GP* model. (b),(e),(h): PAMs displaying contribution of each feature index for DEAP, BVHP, Stroop datasets for the *Feat-GP* model. (c),(f),(i): PAMs displaying the contribution of each data patch for DEAP, BVHP, Stroop datasets for the *PhGP* model.

in the DEAP dataset (up to 12% improvement), while the opposite occurs in the BVHP and Stroop datasets. Overall, *PhGP* improves on the LOSO accuracy obtained by the *SVM-RFE* model by 17%, 5% and 12%, respectively, for the DEAP, BVHP and Stroop datasets.

Note that the results in Table I are obtained by using the full EDA prior model, that is, consisting of all the indices, $\omega_{EDA}(\mathbf{s})$, discussed in Section III-D. In Figure 3 we instead investigate the effect that each subset of features comprising $\omega_{EDA}(\mathbf{s})$, namely, $\omega_{r,p}(\mathbf{s})$, $\omega_t(\mathbf{s})$, $\omega_x(\mathbf{s})$, has on the performance of the *PhGP* model.

We observe that choosing the full set of features (i.e., ω_{EDA} and blue bars in the plots) obtains the highest balanced performance between LOSO sensitivity and specificity of the prediction and therefore highest accuracy compared to when selecting subsets of features for all datasets. Although the feature subsets with ω_x and ω_t vectors result in higher sensitivity than ω_{EDA} in the DEAP and BVHP datasets, the specificity is very low (53% for ω_x and 58% for ω_t in the DEAP dataset and 56% for ω_x and 66% for ω_t in the BVHP dataset). This trend is different in the Stroop dataset, where the sensitivity and specificity is the highest considering the ω_{EDA} feature set.

B. Interpretability Results

We apply the method described in Section III-C to perform interpretability analysis of the trained *PhGP* model on the three datasets for each subject. For simplicity, we focus on the linear prior model (which has better overall accuracy for *PhGP*), though similar results can be obtained using the polynomial and the trigonometric prior by employing techniques discussed in the Supplementary Materials. We also apply the methods from [35], [36] to analyse the interpretability of the trained *Raw-GP* and *Feat-GP* models on each dataset.

We present the physiological activation maps (PAM) derived for each GP-based recognition model and briefly discuss how these maps can help in gaining better understanding of the trained model. Figure 4 shows the generated PAM maps for *Raw-GP*, *Feat-GP* and *PhGP* models trained on the three datasets.

The PAM heatmaps show the normalized value (between zero and one) of the interpretability metric (ϕ, refer to section III-C) for each subject. The vertical axis represents the subject index, which, given the analysis is done in LOSO settings, means a different GP, i.e., one that was learned from the remaining training data. The horizontal axes in the first and third columns (corresponding to *Raw-GP* and *PhGP* models)

TABLE I: Recognition results of the *Raw-GP*, *Feat-GP* and *PhGP* models (considering different forms of functions for parametric modeling of the prior distribution) and the *SVM-RFE* algorithm for DEAP, BVHP and Stroop datasets. The values from left to right are expressed as percentages of accuracy, sensitivity and specificity of the performance of the model.

GP prior	***Raw-GP***	***Feat-GP***	***PhGP***	***SVM-RFE***
Dataset DEAP Acc.(sens., spec.) %				
Zero	65 (84, 47)	64 (74, 56)	-	-
Const.	65 (67, 63)	64 (72, 52)	-	-
Lin.	68 (69, 67)	78 (69, 87)	81 (81, 82)	-
Quad.	72 (74, 70)	78 (77, 80)	73 (68, 78)	-
Cubic	69 (76, 62)	68 (77, 57)	73 (89, 62)	-
Trig.	70 (73, 68)	65 (73, 58)	72 (92, 51)	-
SVM RFE	-	-	-	64 (60, 68)
Dataset BVHP Acc.(sens., spec.) %				
Zero	88 (87, 88)	86 (85, 86)	-	-
Const.	88 (87, 88)	87 (85, 90)	-	-
Lin.	86 (93, 78)	86 (85, 86)	88 (87, 89)	-
Quad.	85 (85, 85)	83 (83, 83)	89 (87, 90)	-
Cubic	89 (90, 87)	83 (84,83)	89 (90, 87)	-
Trig.	88 (87, 89)	86 (84, 89)	88 (87, 89)	-
SVM RFE	-	-	-	84 (81, 87)
Dataset Stroop Acc.(sens., spec.) %				
Zero	85 (85, 85)	80 (88, 73)	-	-
Const.	86 (85, 88)	82 (88, 76)	-	-
Lin.	88 (88, 88)	77 (85, 70)	89 (88, 91)	-
Quad.	89 (88, 91)	80 (82, 79)	83 (85, 82)	-
Cubic	88 (88, 88)	83 (85, 82)	86 (91, 82)	-
Trig.	83 (82, 85)	82 (88, 76)	82 (78, 85)	-
SVM RFE	-	-	-	77 (72, 82)

represent the ϕ value for each selected data patch from the raw EDA signal, whereas in the second column (i.e., the *Feat-GP* model), ϕ is obtained for each feature index in the $\omega_{EDA}(\mathrm{s})$ vector. In all the PAMs the color-bar varies from blue (denoting the value 0 for ϕ) to red (showing the highest obtained value for ϕ). Therefore, the blue blocks represent the lowest contribution of the data patch/feature index for a particular subject in the recognition model, whereas the highest contributions are indicated by the red-colored blocks in the heatmaps.

For all the models we consider 10 values for γ, equally distanced between 0 and 1. Note that the x-axis for *Raw-GP* and *PhGP* models represents time, whereas for the *Feat-GP* it indicates feature indices and the order in which they are shown is arbitrary.

The key observations from the panels (a, c, d, f, g, i) in PAMs (Figure 4) are the patches corresponding to the highest contribution (red color) in the prediction outcome, which are possibly the salient regions in the signal corresponding to the physiological alternations. On the other hand, the panels (b, e, h) indicate the most important features extracted from the EDA signal that capture the changes in psycho-physiological condition of subjects.

From the PAMs of the *Raw-GP* model in the DEAP dataset, we observe that the patches corresponding to the 48th-52nd seconds of the whole 60s duration of the EDA acquisition account for the highest contribution in recognition. On the other hand, the first 18 seconds and the 36th-42nd seconds of the data show the least contribution in almost all subjects. This trend is different in the BVHP dataset, where the highest contribution corresponds to the 2nd and the 8th seconds. On the other hand, the patches occurring around the 2th-3th second and the 5th second are the patches corresponding to the highest contribution in the final prediction in the Stroop dataset.

The PAMs corresponding to the *Feat-GP* model show the high contribution of the $SumAmpSCR$ index (refer to Section 7) in the DEAP dataset. Although the same feature has a relatively high contribution in the BVHP dataset, the highest contribution is obtained through the $Tonicstd$ (refer to Section 7). Similarly, both phasic and tonic related indices show a high contribution in the final prediction in the Stroop dataset.

Concerning the *PhGP* model, the patches of data corresponding to the highest value of ϕ are located at the 18th-26th and 44th-51st seconds in the DEAP dataset and at the 5th and 2nd seconds for the BVHP and Stroop datasets, respectively.

VI. Discussions

In this study we presented a novel approach for designing an interpretable recognition model using Bayesian GP classification. We proposed two levels of interpretability: i) at a model design level, we have proposed *PhGP* modelling which is more transparent compared to traditional GP models, through embedding physiologically-based mathematical models within the GP inference. This level of interpretability is more acceptable for the clinicians since their domain knowledge is taken into account. ii) at a post-hoc level, thanks to the analytical formulation of *PhGP*, it is amenable to interpretability analysis with the methods discussed in Section III-C.

The main difference between our approach and existing feature-based methods lies in the way we explicitly inject expert knowledge into the learning algorithm of the ML model in the form of previously validated physiologically-inspired models and assumptions. While feature-based approaches may utilise such expert knowledge by considering the hand-crafted features in the input space, they do not inform the learning procedure of the ML model about it. Furthermore notice that in *PhGP*, the interpretability metric we provide is computed formally, with provable bounds and not approximated with gradient techniques. It is important to note that the innovation in the *PhGP* design provides interesting insights into identifying the salient regions in the input, in view of access to both the raw physiological signal and physiologically-based features.

Comparison of our *PhGP* model with the *Raw-GP*, *Feat-GP* and *SVM-RFE* models in Table I demonstrate the merit of relying in recognition tasks on physiological signal analysis and information from end-to-end modelling, i.e., *PhGP*, by drawing on both aspects, is able to achieve competitive performance in all datasets. Moreover, although potentially having access to the same information (that is, the full raw signal), the *Raw-GP* model tends to overfit, while the *PhGP* methods benefit from the physiologically-informed prior in shaping its output distribution.

It is interesting to note how all the GP-based models outperform the SVM-RFE method; in fact, the latter tends to overfit in these settings. Furthermore, by using MLE for the hyper-parameters of the prior in the GP settings, we also obtain a form of feature selection in the prior space, though in an approximate Bayesian fashion, which provides better generalisation properties. The *PhGP* model offers higher accuracy compared to a recent study performed in similar settings on the DEAP dataset, which obtained 71% accuracy [19] (similar to that obtained by SVM-RFE). As in here, previous studies have conducted experiments on the BVHP dataset with the aim of classifying the lowest level of pain from the highest pain threshold level and validated their results with LOSO cross-validation, achieving 77% [39] and 79% [38] of accuracy. Interestingly, *PhGP* improves on the accuracy of all these methods, although it targets a more difficult task of recognition, that is, classifying between the minimum and the maximum level of the pain stimulus.

PhGP obtains comparative results both to discriminate low vs. high sympathetic discharge as in BVHP and Stroop datasets, and when the sympathetic activity is similarly triggered by two different emotional processes as in the DEAP dataset. This latter aspect suggests how sympathetic activity is not a monotonous and stereotypical reaction but is modulated by the activating stimulus.

We highlight that, in addition to obtaining good accuracy performance of GP-based models, they provide different insights into the salient regions in the input data which possibly correspond to the patches in the input data where the highest alterations in physiology are present. While the PAMs in Figure 4 obtained from *Raw-GP* and *Feat-GP* indicate the most informative regions in the raw signal and the most important features, respectively, both sources of information are inherently reflected in the salient regions obtained from *PhGP*.

It is relevant to highlight how the raw signal and the features which are more informative are different and specific for each dataset. The three datasets reflect three different triggers for sympathetic response: physical stressor (i.e., pain); cognitive stressor (i.e., stroop); emotional stressor (i.e., emotional pictures). Once again this might suggest that sympathetic response, as measured with EDA, is not monotonous and stereotyped but also depends on the nature of the stressor. Further studies are needed to specifically test this hypothesis and to understand the role of this sympathetic specific response.

Moreover, it is interesting to observe the consistency in the ϕ values reported in each of PAMs in Figure 4. Those refer to different models learnt in the same settings (only the training/test set split varies), so this highlights how the results obtained by interpretability analysis are in a sense qualitatively independent from the specific subjects. This indicates that the model is learning features and patterns that are specific to the problem itself, rather than the particular subject involved.

Considering the EDA trend depicted in Figure 2, it is evident that the EDA dynamics is different for each psycho-physiological condition. In the DEAP dataset, the EDA dynamics in a positive valence condition shows an elevated response from the 45th second to the end of the stimulation, while the negative valence condition is relatively smoother along the timeline with a peak response at the 18th second. For pain stimulation, while during the lowest level of pain a decreasing trend in EDA response is observed, during the highest level of pain a maximum peak of response is observed at the 8th second of the stimulation. During the Stroop task, a higher elevation in EDA response is observed compared to the resting state where the fluctuations are relatively lower in amplitude.

Although visualization of these plots can aid understanding of the different trends in each psycho-physiological condition, quantifying this difference is a difficult task. The PAMs obtained as a result of interpretability analysis introduced in this paper are a first step to quantify the position of those patches where the highest difference in the EDA response between the two psycho-physiological conditions is obtained.

VII. Conclusions

In this paper we provided an interpretable GP-based framework that facilitates the injection of physiologically-inspired priors in a Bayesian GP model in order to infer a user's psycho-physiological state. The experimental application of the proposed *PhGP* model on EDA signals in this study indicates that *PhGP* not only obtains comparative performance among competitive predictive models, but also provides physiologically sound interpretation of predictions which are consistent at the single-subject level. This generality of the results is a remarkable feature of the *PhGP* model for recognition tasks. We believe that our methodology offers considerable advantages for recognition systems that input signals from non-invasive wearable monitoring systems (e.g, smart-watches, sensorised gloves or shirts), where recording EDA is easy and inexpensive. From a clinical perspective, the proposed method will be able to support the clinician, operating for example in the psychological field, by providing not only an automatic diagnosis based on objective measures such as those of physiological data but also a tool capable of helping to understand the psycho-physiological motivation behind this diagnosis.

Although the experimental applications in this study are limited to the EDA signal, the modelling can be adapted to other available models for the analysis of physiological signals (e.g., *point-process* modelling of heartbeat dynamics and *recursive penalized least squares solution* for EEG generation) in the *PhGP* model. An extension of the current methodology to a multi-modal *PhGP* model that benefits from meaningful information content of different physiological signals is a focus for future work. The methodology for interpretability analysis of the *PhGP* model in this paper provides useful insights about the predictions made by the model, but this is just an initial step. In future we aim to provide systematic means of interpretability to investigate more complex properties of psycho-physiological mechanisms such as correlational and causal relationships. Moreover, we will explore methods to combine deep neural networks with GPs to encode the prior physiological model while keeping the interepretability of the framework.

References

[1] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[2] R. W. Picard, *Affective computing*. MIT press, 2000.

[3] M. A. Ahmad, C. Eckert, and A. Teredesai, "Interpretable machine learning in healthcare," in *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, 2018, pp. 559–560.

[4] X. Zhou, K. Jin, Y. Shang, and G. Guo, "Visually interpretable representation learning for depression recognition from facial images," *IEEE Transactions on Affective Computing*, 2018.

[5] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015.

[6] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, 2018.

[7] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[8] A. Patane, S. Ghiasi, E. P. Scilingo, and M. Kwiatkowska, "Automated recognition of sleep arousal using multimodal and personalized deep ensembles of neural networks," in *2018 Computing in Cardiology Conference (CinC)*, vol. 45. IEEE, 2018, pp. 1–4.

[9] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[10] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "cvxeda: A convex optimization approach to electrodermal activity processing," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. 797–804, 2016.

[11] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[12] S. Walter, S. Gruss, H. Ehleiter, J. Tan, H. C. Traue, P. Werner, A. Al-Hamadi, S. Crawcour, A. O. Andrade, and G. M. da Silva, "The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system," in *2013 IEEE international conference on cybernetics (CYBCO)*. IEEE, 2013, pp. 128–131.

[13] S. Ghiasi, A. Patane, A. Greco, L. Laurenti, E. Scilingo, and M. Kwiatkowska, "Gaussian processes with physiologically-inspired priors for physical arousal recognition," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 54–57.

[14] J. Mateo and P. Laguna, "Improved heart rate variability signal analysis from the beat occurrence times according to the ipfm model," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 8, pp. 985–996, 2000.

[15] G. Valenza, L. Citi, E. P. Scilingo, and R. Barbieri, "Point-process nonlinear models with laguerre and volterra expansions: Instantaneous assessment of heartbeat dynamics," *IEEE Transactions on Signal Processing*, vol. 61, no. 11, pp. 2914–2926, 2013.

[16] D. R. Bach and K. J. Friston, "Model-based analysis of skin conductance responses: Towards causal models in psychophysiology," *Psychophysiology*, vol. 50, no. 1, pp. 15–22, 2013.

[17] O. Yamashita, A. Galka, T. Ozaki, R. Biscay, and P. Valdes-Sosa, "Recursive penalized least squares solution for dynamical inverse problems of eeg generation," *Human brain mapping*, vol. 21, no. 4, pp. 221–235, 2004.

[18] P. V. Rouast, M. Adam, and R. Chiong, "Deep learning for human affect recognition: insights and new developments," *IEEE Transactions on Affective Computing*, 2019.

[19] S. Siddharth, T.-P. Jung, and T. J. Sejnowski, "Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing," *IEEE Transactions on Affective Computing*, 2019.

[20] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[21] A. Patane and M. Kwiatkowska, "Calibrating the classifier: siamese neural network architecture for end-to-end arousal recognition from ecg," in *International Conference on Machine Learning, Optimization, and Data Science*. Springer, 2018, pp. 1–13.

[22] J. Huang, Y. Li, J. Tao, Z. Lian, M. Niu, and M. Yang, "Multimodal continuous emotion recognition with data augmentation using recurrent neural networks," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, 2018, pp. 57–64.

[23] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Progressive neural networks for transfer learning in emotion recognition," *arXiv preprint arXiv:1706.03256*, 2017.

[24] N. Jaques, S. Taylor, A. Sano, R. Picard *et al.*, "Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation," in *IJCAI 2017 Workshop on artificial intelligence in affective computing*, 2017, pp. 17–33.

[25] J. Wilzen, A. Eklund, and M. Villani, "Physiological gaussian process priors for the hemodynamics in fmri analysis," *arXiv preprint arXiv:1708.06152*, 2017.

[26] e. a. Clifton, Lei, "Gaussian process regression in vital-sign early warning systems," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2012, pp. 6161–6164.

[27] S. M. Imaduddin, A. Fanelli, F. Vonberg, R. C. Tasker, and T. Heldt, "Pseudo-bayesian model-based noninvasive intracranial pressure estimation and tracking," *IEEE Transactions on Biomedical Engineering*, 2019.

[28] H. F. García, M. A. Álvarez, and Á. A. Orozco, "Gaussian process dynamical models for multimodal affect recognition," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 850–853.

[29] J. Quinonero-Candela, "Learning with uncertainty: Gaussian processes and relevance vector machines," Ph.D. dissertation, Technical University of Denmark Lyngby, Denmark, 2004.

[30] L. Martino and J. Read, "A joint introduction to gaussian processes and relevance vector machines with connections to kalman filtering and other kernel smoothers," *Information Fusion*, vol. 74, pp. 17–38, 2021.

[31] A. Greco, G. Valenza, L. Citi, and E. P. Scilingo, "Arousal and valence recognition of affective sounds based on electrodermal activity," *IEEE Sensors Journal*, vol. 17, no. 3, pp. 716–725, 2016.

[32] P. Werner, D. Lopez-Martinez, S. Walter, A. Al-Hamadi, S. Gruss, and R. Picard, "Automatic recognition methods supporting pain assessment: A survey," *IEEE Transactions on Affective Computing*, 2019.

[33] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer School on Machine Learning*. Springer, 2003, pp. 63–71.

[34] C. E. Rasmussen and H. Nickisch, "Gaussian processes for machine learning (gpml) toolbox," *Journal of machine learning research*, vol. 11, no. Nov, pp. 3011–3015, 2010.

[35] A. Blaas, A. Patane, L. Laurenti, L. Cardelli, M. Kwiatkowska, and S. Roberts, "Adversarial robustness guarantees for classification with gaussian processes," in *International Conference on Artificial Intelligence and Statistics*, 2020, pp. 3372–3382.

[36] A. Patane, A. Blaas, L. Laurenti, L. Cardelli, S. Roberts, and M. Kwiatkowska, "Adversarial robustness guarantees for gaussian processes," *arXiv preprint arXiv:2104.03180*, 2021.

[37] S. Ghiasi, A. Greco, R. Barbieri, E. P. Scilingo, and G. Valenza, "Assessing autonomic function from electrodermal activity and heart rate variability during cold-pressor test and emotional challenge," *Scientific Reports*, vol. 10, no. 1, pp. 1–13, 2020.

[38] M. Kächele, P. Thiam, M. Amirian, P. Werner, S. Walter, F. Schwenker, and G. Palm, "Multimodal data fusion for person-independent, continuous estimation of pain intensity," in *International Conference on Engineering Applications of Neural Networks*. Springer, 2015, pp. 275–285.

[39] P. Thiam, P. Bellmann, H. A. Kestler, and F. Schwenker, "Exploring deep physiological models for nociceptive pain recognition," *Sensors*, vol. 19, no. 20, p. 4503, 2019.

[40] F. Scarpina and S. Tagini, "The stroop color and word test," *Frontiers in psychology*, vol. 8, p. 557, 2017.

[41] K. Yan and D. Zhang, "Feature selection and analysis on correlated gas sensor data with recursive feature elimination," *Sensors and Actuators B: Chemical*, vol. 212, pp. 353–363, 2015.

I. Supplementary materials (methodology)

A. Classification with GP Models

The first step in GP classification is that of putting a GP prior over the latent function variable f, that is, for $x \in \mathbb{R}^n$ we assume that

$$f(x) \sim p(f|x) = \mathcal{N}(f; \mu(x), k(x,x)),$$

for a specific choice of the mean function μ and kernel (or covariance) function k. Generally, the mean function and the kernel function are controlled by a set of hyper-parameters, which we denote respectively with α and β. The GP prior induces a multi-variate Gaussian prior distribution on the vector of latent functions over the training set, $\mathbf{f} = [f(x^{(1)}), \ldots, f(x^{(N)})]$, i.e. $p(\mathbf{f}|\mathbf{x})$. Learning in Bayesian settings amounts to the computation of the posterior distribution $p(\mathbf{f}|\mathcal{D})$ via the Bayes formula.

Given a test point $x^* \in \mathbb{R}^n$, estimation of the output on x^* is obtained by computing its posterior latent distribution $p(f^*|\mathcal{D}, x^*)$, for $f^* \in \mathbb{R}$, which is then integrated for the probability that x^* belongs to class 1 as follows:

$$p(f^*|\mathcal{D}, x^*) = \int p(f^*|\mathbf{x}, x^*, \mathbf{f}) p(\mathbf{f}|\mathcal{D}) d\mathbf{f}$$

$$p(y=1|\mathcal{D}, x^*) = \int \sigma(f^*) p(f^*|\mathcal{D}, x^*) df^*.$$

Unfortunately, because of the non-Gaussian nature of the likelihood function, the integrals above cannot be computed analytically [1]. We rely on the Laplace method for GP classification inference, which proceeds by computing a Gaussian approximation $q(f^*|\hat{\mu}(x^*), \hat{\Sigma}(x^*)) = \mathcal{N}(f^*|\hat{\mu}(x^*), \hat{\Sigma}(x^*))$ of the latent posterior distribution. Specifically, let $\mathbf{k}^* = [k(x^*, x^{(1)}), \ldots, k(x^*, x^{(N)})] \in \mathbb{R}^N$ be the vector of co-variances between the test point x^* and the points in the dataset $\mathcal{D}$ and let $K \in \mathbb{R}^{N \times N}$ be the matrix of covariances between the training points. Then the Laplace approximate posterior mean and variance are:

$$\hat{\mu}(x^*) = \mu(x^*) + \mathbf{k}^{*T} K^{-1} \hat{\mathbf{f}} \tag{1}$$

$$\hat{\Sigma}(x^*) = k(x^*, x^*) - \mathbf{k}^{*T} (K + W)^{-1} \mathbf{k}^* \tag{2}$$

where $\hat{\mathbf{f}}$ is the mode of $p(\mathbf{f}|\mathcal{D})$ and W is the Hessian of the negative log-likelihood. Note how the prior mean and variance are adjusted in the learning process according to the information contained in the training data and the kernel scale captured by the hyper-parameters.

Finally, given a vector of latent variables over the training set, $\mathbf{f} = [f^{(1)}, \ldots, f^{(N)}] \in \mathbb{R}^N$, its likelihood with respect to $\mathcal{D}$ can be computed as:

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{N} \left[\sigma(f^{(i)})^{y^{(i)}} (1 - \sigma(f^{(i)}))^{1-y^{(i)}} \right], \tag{3}$$

which represents the likelihood of observing the class vector $\mathbf{y}$ given specific values for the latent variable vector $\mathbf{f}$.

B. Hyper-parameter Estimation for PhGP

In order to estimate the hyper-parameters α and β on the mean and kernel function in the case of *PhGP* we adapt the maximum likelihood framework (MLE), and show how it can be employed straightforwardly on a MAP approximation of the physiological model.

In fact, by marginalising the latent variable from the GP likelihood (Equation (3)) and applying the inference formulas of the Laplace approximation, we obtain the marginal log-likelihood as:

$$\log p(\mathbf{y}|\mathcal{D}, \alpha, \beta) = \log \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathcal{D}, \alpha, \beta) d\mathbf{f} =$$
$$-\frac{1}{2} \hat{\mathbf{f}}^T K^{-1} \hat{\mathbf{f}} + \log p(\mathbf{y}|\mathbf{f}) - \frac{1}{2} \log |I + W^{\frac{1}{2}} K W^{\frac{1}{2}}|, \tag{4}$$

where K explicitly depends on β, while both $\hat{\mathbf{f}}$ and W implicitly depend on α and β. Equation (4) can be optimised for the values of the hyper-parameters that best justify the training data $\mathcal{D}$, which can be shown to provide an approximation of the MLE for α and β [1]. To achieve this in the case of *PhGP*, we can apply standard gradient-based optimisation method typically used for GP models, by additionally propagating the derivatives wrt α through Equation (4) in the main text. To do so, we proceed by approximating the derivative computation by considering the MAP solution for the sub-process $\mathbf{s}$, so that we have that $\frac{d\mu(x|\alpha)}{d\alpha} \approx \frac{dm_j(\mathbf{s}_{\text{MAP}}|\alpha)}{d\alpha}$. The latter is straightforward to compute as $\mathbf{s}_{\text{MAP}}$ does not depend on any hyper-parameters.

C. Interpretability Framework of PhGP Models

To compute the quantities $\phi(T^I_{\gamma,x*})$ and $\Phi(\mathcal{T}, I, \gamma)$ defined in Definition 1 of the main text, it suffices to compute the minimum and maximum classification probabilities for $x \in T^I_{\gamma,x^*}$. This problem has been studied for GPs in [2], where it has been shown that it reduces to the computation of the minimum and maximum of $\hat{\mu}(x)$ and $\hat{\Sigma}(x)$ for $x \in T^I_{\gamma,x^*}$, that is, of the posterior mean and variance. In order to generalise the bounds in the case of the *PhGP* model, it suffices to additionally compute lower and upper bounds over the a-priori mean function, that is, $\mu^{L,\text{pr}}_T$ and $\mu^{U,\text{pr}}_T$ such that:

$$\mu^{L,\text{pr}}_T \leq \min_{x \in T} \mu(x|\alpha) \qquad \mu^{U,\text{pr}}_T \geq \max_{x \in T} \mu(x|\alpha).$$

How to compute suitable values for $\mu^{L,\text{pr}}_T$ and $\mu^{U,\text{pr}}_T$ is a problem that depends, of course, on the exact form of the prior function used. For prior mean functions that can be written down analytically, lower and upper bounds can be computed by relying on interval bound propagation techniques, though in the general case one might have to resort to numerical optimisation methods if smoothness assumptions are not satisfied. The bounding problem is actually quite simple for the polynomial and trigonometric functions introduced in Equations (5) and (6).

$$m_1(\mathbf{s}|\alpha) = \sum_{p=1}^{d} \sum_{j=1}^{m} \alpha_{pj} \omega_j(\mathbf{s})^p \tag{5}$$

$$m_2(\mathbf{s}|\alpha) = \sum_{p=1}^{d} \alpha^{(1)}_p \cos \left(\sum_{j=1}^{m} \alpha^{(2)}_{pj} \omega_j(\mathbf{s}) + \alpha^{(3)}_p \right), \tag{6}$$

In particular, in the polynomial case the overall solution can be written down in closed from, as stated in the following proposition.

Proposition 1: For $j \in \{1, ..., m\}$ and $p \in \{1, ..., d\}$, let $T = [\omega^L, \omega^U] \subset \mathbb{R}^m$ be a hyper-rectangle in the feature space such that:

$$\omega_j^L \leq \omega_j(\mathbf{s}(x)) \leq \omega_j^U \qquad \forall x \in T^I_{\gamma, x^*}. \tag{7}$$

Let $\omega_j^{L,p} = \min_{\omega \in \{\omega_j^L, \omega_j^U\}} \omega_j^p$, $\omega_j^{U,p} = \max_{\omega \in \{\omega_j^L, \omega_j^U\}} \omega_j^p$ and

$$\{\bar{\omega}_j^{L,p}, \bar{\omega}_j^{U,p}\} = \begin{cases} \{\omega_j^{L,p}, \omega_j^{U,p}\} & \text{if } \alpha_{p,j} \geq 0 \\ \{\omega_j^{U,p}, \omega_j^{L,p}\} & \text{otherwise} \end{cases}.$$

Then, it holds that

$$\sum_{p=1}^{d} \sum_{j=1}^{m} \alpha_{p,j} \bar{\omega}_j^{l,p} \leq \mu(x|\alpha) \leq \sum_{p=1}^{d} \sum_{j=1}^{m} \alpha_{p,j} \bar{\omega}_j^{u,p}. \tag{8}$$

For the case of the trigonometric prior function, similar bounds can be obtained by further accounting for the periodicity of the cosine function in the computation. Notice that the above proposition relies on the computation of upper and lower bounds on the feature vector given a range on the raw signal (see Equation (7)). In general, this cannot be computed exactly because the integral in Equation (4) of the main text is intractable. As for the hyper-parameters, we can rely on a MAP estimation of the processes $\mathbf{s}$, which gives us a pointwise estimate for the feature vector ω. It is then easy to see how rectangular bounds on the input space x can be propagated in a straightforward fashion for time-domain and statistical features such as mean, standard deviation and min/max of the signals. For the general case, e.g., frequency-domain features, we instead need to rely on numerical optimisation methods for the approximation of the bounds.

Using the above proposition, in conjunction with the methods presented in [2], [3], it is possible to estimate the interpretability metric of Definition 1 for the *PhGP* model, and further refine the approximations by means of a branch-and-bound technique.

In fact, as the prior mean has only an additive effect on the posterior computation, an evaluation of the posterior effect induced from the dataset proceeds similarly to how this is done for standard classification GPs.

D. SVM-RFE algorithm

We applied standard nonlinear Support Vector Machine (SVM) embedded with Recursive Feature Elimination (RFE) on the aforementioned experimental data. We chose this method for comparison with GP-based models since it has been widely accepted as one of the best classification methods in terms of performance and interpretability [4]. RFE is an embedded feature selection method based on a backward sequential selection that eliminates a feature in a feature set of size m that has the least effect on the SVM weight-vector norm at each iteration. This way, the features are ranked and the SVM classification is repeated m times while the last ranked features are removed. Finally, a subset of features with size r that optimises the performance of the SVM classifier are selected. We used the open source toolkit LIBSVM for the implementation of SVM [5].

TABLE I: Recognition results of 1D-CNN model for DEAP, BVHP and Stroop datasets.

Dataset	Sensitivity	Specificity	Accuracy	AUCROC	F-score
Deap	65	74	70	0.68	71
BVHP	83	87	85	85	84
Stroop	85	78	82	79	80

TABLE II: Recognition results using K-means cross validation.

Recognition Model	DEAP			BVHP			Stroop		
	Sens.	Spec.	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.	Acc.
Raw-GP	65 ± 2	68 ± 5	67 ± 3	94 ± 7	73 ± 5	84 ± 6	86 ± 4	87 ± 4	87 ± 4
Feat-GP	72 ± 6	88 ± 4	80 ± 5	83 ± 2	88 ± 4	86 ± 3	83 ± 3	75 ± 2	79 ± 2
PhGP	83 ± 2	85 ± 4	84 ± 3	88 ± 3	89 ± 3	89 ± 3	88 ± 3	91 ± 3	89 ± 3
SVM-RFE	63 ± 6	65 ± 3	64 ± 4	78 ± 4	85 ± 2	82 ± 3	70 ± 4	84 ± 5	77 ± 4

II. Supplementary materials (experiments)

A. Comparison with deep neural network (DNN) models

We have applied 1D-CNN model as in order to compare our results with baseline DNN based models. The results are reported in Table I. We constructed this DNN model using convolutional layer, max pooling layer and a fully connected Softmax layer for the classification of electrodermal activity signals. We have used ADAM algorithm for the optimization process. The results are reported within the LOSO cross validation scheme.

B. Recognition results using K-means cross validation.

We report the performance metrics within the K-means cross validation (K=5) considering the linear mean function for the GP-based models (Table II).

C. Computational time

We have compared the computation speed of the three GP based in Table III for each single epoch and the total number of epochs used to obtain the reported results.

TABLE III: Comparison of the computational time of the GP based models. Values are presented as the computational time per single epoch in seconds

Recognition Model	DEAP	BVHP	Stroop
Raw-GP	0.63	0.99	0.32
Feat-GP	0.56	0.57	0.79
PhGP	6.54	10.20	3.53

References

[1] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer School on Machine Learning*. Springer, 2003, pp. 63–71.

[2] A. Blaas, A. Patane, L. Laurenti, L. Cardelli, M. Kwiatkowska, and S. Roberts, "Adversarial robustness guarantees for classification with gaussian processes," in *International Conference on Artificial Intelligence and Statistics*, 2020, pp. 3372–3382.

[3] A. Patane, A. Blaas, L. Laurenti, L. Cardelli, S. Roberts, and M. Kwiatkowska, "Adversarial robustness guarantees for gaussian processes," *arXiv preprint arXiv:2104.03180*, 2021.

[4] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, 2018.

[5] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.