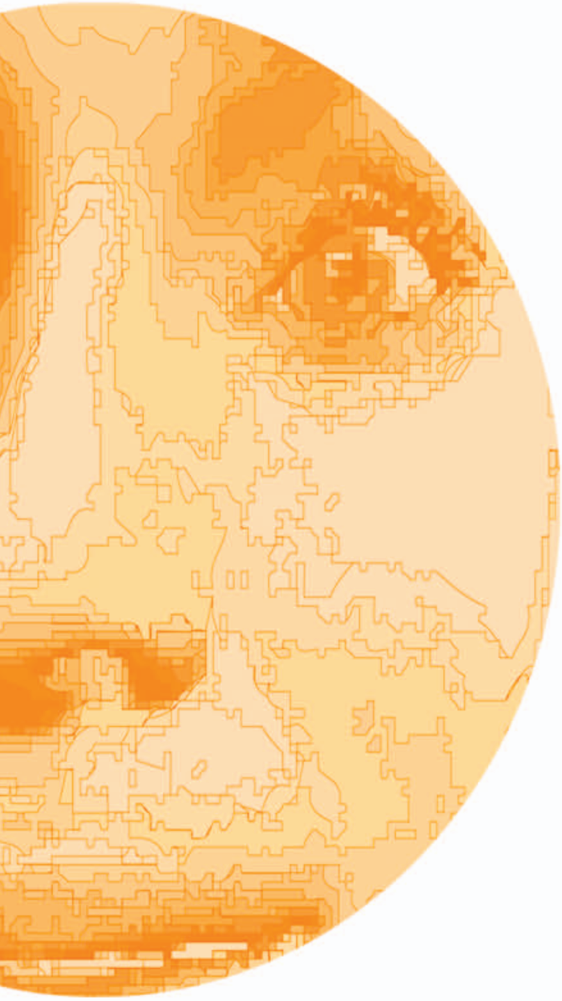
Learning Deep Physiological Models of Affect

© PHOTODISC

Héctor P. Martínez
IT University of Copenhagen, DENMARK

Yoshua Bengio
University of Montreal, CANADA

Georgios N. Yannakakis
University of Malta, MALTA



I. Introduction

More than 15 years after the early studies in *Affective Computing* (AC), [1] the problem of detecting and modeling emotions in the context of human-computer interaction (HCI) remains complex and largely unexplored. The detection and modeling of emotion is, primarily, the study and use of artificial intelligence (AI) techniques for the construction of computational models of emotion. The key challenges one faces when attempting to model emotion [2] are inherent in the vague definitions and fuzzy boundaries of emotion, and in the modeling methodology followed. In this context, open research questions are still present in all key components of the modeling process. These include, first, the appropriateness of the modeling tool employed to map emotional manifestations and responses to annotated affective states; second, the processing of signals that express these manifestations (i.e., model input); and third, the way affective annotation (i.e., model output) is handled. This paper touches upon all three key components of an affective model (i.e., input, model, output) and introduces the use of *deep learning* (DL) [3], [4], [5] methodologies for affective modeling from multiple physiological signals.

Traditionally in AC research, behavioral and bodily responses to stimuli are collected and used as the affective model input. The input can be of three main types: a) behavioral responses to emotional stimuli expressed through an

Digital Object Identifier 10.1109/MCI.2013.2247823
Date of publication: 11 April 2013

1556-603X/13/$31.00©2013IEEE

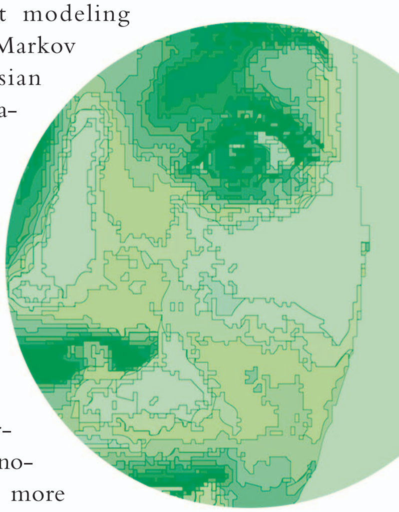
Abstract**—**Feature extraction and feature selection are crucial phases in the process of affective modeling. Both, however, incorporate substantial limitations that hinder the development of reliable and accurate models of affect. For the purpose of modeling affect manifested through physiology, this paper builds on recent advances in machine learning with deep learning (DL) approaches. The efficiency of DL algorithms that train artificial neural network models is tested and compared against standard feature extraction and selection approaches followed in the literature. Results on a game data corpus—containing players' physiological signals (i.e., skin conductance and blood volume pulse) and subjective self-reports of affect—reveal that DL outperforms manual ad-hoc feature extraction as it yields significantly more accurate affective models. Moreover, it appears that DL meets and even outperforms affective models that are boosted by automatic feature selection, for several of the scenarios examined. As the DL method is generic and applicable to any affective modeling task, the key findings of the paper suggest that ad-hoc feature extraction and selection—to a lesser degree—could be bypassed.

interactive application (e.g., data obtained from a log of actions performed in a game); b) objective data collected as bodily responses to stimuli, such as physiological signals and facial expressions; and c) the context of the interaction. Before these data streams are fed into the computational model, an automatic or ad-hoc *feature extraction* procedure is employed to derive appropriate signal attributes (e.g., average skin conductance) that will feed the model. It is also common to introduce an automatic or a semi-automatic *feature selection* procedure that picks the most appropriate of the features extracted.

While the phases of feature extraction and feature selection are beneficial for affective modeling, they inherit a number of critical limitations that make their use cumbersome in highly complex multimodal input spaces. First, manual feature extraction limits the creativity of attribute design to the expert (i.e., the AC researcher) resulting in potentially inappropriate affect detectors that might not be able to capture the manifestations of the affect embedded in the raw input signals. Second, both feature extraction and feature selection—to a larger degree—are computationally expensive phases. In particular, the computational cost of feature selection may increase combinatorially (quadratically, in the greedy case) with respect to the number of features considered [6]. In general, there is no guarantee that any search algorithm is able to converge to optimal feature sets for the model; even exhaustive search may be approximate, since models are often trained with non-deterministic algorithms.

Our hypothesis is that the use of non-linear unsupervised and supervised learning methods relying on the principles of DL [3], [4] can eliminate the limitations of the current feature extraction and feature selection practices in affective modeling. We test the hypothesis that DL could construct feature extractors that are more appropriate than selected adhoc features picked via automatic selection. Learning within deep artificial neural network (ANN) architectures has proven to be a powerful machine learning approach for a number of benchmark problems and domains, including image and speech recognition [7], [8]. DL allows the automation of feature extraction (and feature selection, in part) without compromising on the accuracy of the obtained computational models and the physical meaning of the data attributes extracted [9]. Using deep learning we were able to extract meaningful multimodal data attributes beyond manual ad-hoc feature design. These learned attributes led to more accurate affective models and, at the same time, potentially save computational resources by bypassing the computationally expensive feature selection phase. Most importantly, with the use of DL we gain simplicity as multiple signals can be fused and fed directly—with limited preprocessing—to the model for training.

Other common automatic feature extraction techniques within AC are principal component analysis (PCA) and Fisher projection. However they are typically applied to a set of features extracted a priori [10] while we apply DL directly to the raw data signals. Moreover, DL techniques can operate with any signal type and are not restricted to discrete signals as, for example, sequential data mining techniques are [11]. Finally, compared to dynamic affect modeling approaches such as Hidden Markov Models and Dynamic Bayesian Networks, DL models are advantageous with respect to their ability to reduce signal resolution across the several layers of their architectures.



This paper focuses on developing DL models of affect using data which are annotated in a ranking format (pairwise preferences). We *emphasize* the benefits of preference-based (or ranking-based) annotations for emotion (e.g., X is more frustrating than Y) as opposed to rating-based annotation [12] (such as the self-assessment manikins [13], a tool to rate levels of arousal and valence in discrete or continuous scales [14]) and introduce the use of DL algorithms for preference learning, namely, preference deep learning (PDL). In this paper, the PDL algorithm proposed is tested on emotional manifestations of *relaxation, anxiety, excitement,* and *fun,* embedded in physiological signals (i.e., skin conductance and blood volume pulse) derived from a game-based user study of 36 participants. The study compares DL against ad-hoc feature extraction on physiological signals, used broadly in the AC literature, showing that DL yields models of equal or significantly higher accuracy when a single signal is used as model input. When the skin conductance and blood volume pulse signals are fused, DL outperforms standard feature extraction across all affective states examined. The supremacy of DL is maintained even when automatic feature selection is employed to improve models built on ad-hoc features; in several affective states the performance of models built on automatically selected ad-hoc features does not surpass or reach the corresponding accuracy of the PDL approach.

This paper advances the state-of-the-art in affective modeling in several ways. First, to the best of the authors' knowledge, this is the first time deep learning is introduced to the domain of psychophysiology, yielding efficient computational models of affect. Second, the paper shows the strength of the method when applied to the fusion of different physiological signals. Third, the paper introduces PDL, i.e., the use of deep ANN architectures trained on ranked (pairwise preference) annotations of affect. Finally, the key findings of the paper show the potential of DL as a mechanism for eliminating manual feature extraction and even, in some occasions, bypassing automatic feature selection for affective modeling.

II. Computational Modeling of Affect

Emotions and affect are mental and bodily processes that can be inferred by a human observer from a combination of contextual, behavioral and physiological cues. Part of the complexity of affect modeling emerges from the challenges of finding objective and measurable signals that carry affective information (e.g., body posture, speech and skin conductance) and designing methodologies to collect and label emotional experiences effectively (e.g., induce specific emotions by exposing participants to a set of images). Although this paper is only concerned with computational aspects of creating physiological detectors of affect, the signals and the affective target values collected shape the modeling task and, thus, influence the efficacy and applicability of dissimilar computational methods. Consequently, this section gives an overview of the field beyond the input modalities and emotion annotation protocols examined in our case study. Furthermore, the studies surveyed are representative of the two principal applications of AI for affect modeling and cover the two key research pillars of this paper: 1) defining feature sets to extract relevant bits of information from objective data signals (i.e., for feature extraction), and 2) creating models that map a feature set into predicted affective states (i.e., for training models of affect).

A. Feature Extraction

In the context of affect detection, we refer to feature extraction as the process of transforming the raw signals captured by the hardware (e.g., a skin conductance sensor, a microphone, or a camera) into a set of inputs suitable for a computational predictor of affect. The most common features extracted from unidimensional continuous signals—i.e. temporal sequences of real values such as blood volume pulse, accelerometer data, or speech—are simple statistical features, such as average and standard deviation values, calculated on the time or frequency domains of the raw or the normalized signals (see [15], [16] among others). More complex feature extractors inspired by signal processing methods have also been proposed by several authors. For instance, Giakoumis et al. [17] proposed features extracted from physiological signals using Legendre and Krawtchouk polynomials while Yannakakis and Hallam [18] used the approximate entropy [19] and the parameters of linear, quadratic and exponential regression models fitted to a heart rate signal. The focus of this paper is on DL methods that can automatically derive feature extractors from the raw data, as opposed to a fixed set of hand-crafted extractors that represent pre-designed statistical features of the signals.

Unidimensional symbolic or discrete signals—i.e., temporal sequences of discrete labels, typically *events* such as clicking a mouse button or blinking an eye—are usually transformed with ad-hoc statistical feature extractors such as counts, similarly to continuous signals. Distinctively, Martínez and Yannakakis [11] used frequent sequence mining methods [20] to find frequent patterns across different discrete modalities, namely gameplay events and discrete physiological events. The count of each pattern was then used as an input feature to an affect detector. This methodology is only applicable to discrete signals: continuous signals must be discretized, which involves a loss of information. To this end, the key advantage of the DL methodology proposed in this paper is that it can handle both discrete and continuous signals; a lossless transformation can convert a discrete signal into a binary continuous signal, which can potentially be fed into a deep network—DL has been successfully applied to classify binary images, e.g., [21].

Affect recognition based on signals with more than one dimension typically boils down to affect recognition from images or videos of body movements, posture or facial expressions. In most studies, a series of relevant points of the face or body are first detected (e.g., right mouth corner and right elbow) and tracked along frames. Second, the tracked points are aggregated into discrete *Action Units* [22], gestures [23] (e.g., lip stretch or head nod) or continuous statistical features (e.g., body contraction index), which are then used to predict the affective state of the user [24]. Both above-mentioned feature extraction steps are, by definition, supervised learning problems as the points to be tracked and action units to be identified have been defined a priori. While these problems have been investigated extensively under the name of facial expression or gesture recognition, we will not survey them broadly as this paper focuses on methods for automatically discovering new or unknown features in an unsupervised manner.

Deep neural network architectures such as convolutional neural networks (CNNs), as a popular technique for object recognition in images [25], have also been applied for facial expression recognition. In [26], CNNs were used to detect predefined features such as eyes and mouth which later were used to detect smiles. Contrary to our work, in that study each of the layers of the CNN was trained independently using backpropagation, i.e., labeled data was available for training each level. More recently, Rifai et al. [27] successfully applied a variant of auto-encoders [21] and convolutional networks, namely Contractive Convolutional Neural Networks, to learn features from images of faces and predict the displayed emotion, breaking the previous state-of-the-art on the Toronto Face Database [28]. The key differences of this paper with that study reside in the nature of the dataset and the method used. While Rifai et al. [27] used a large dataset (over 100,000 samples; 4,178 of them were labeled with an emotion class) of static images displaying posed emotions, we use a small dataset (224 samples, labeled with pairwise orders) with a set of physiological time-series recorded along an emotional experience. The reduced size of our dataset (which is of the same magnitude as datasets used in related psycho-physiological studies—e.g., [29], [30]) does not allow the extraction of large feature sets (e.g., 9,000 features in [27]), which would lead to affect models of poor generalizability. The nature of our preference labels also calls for a modified CNN training algorithm for affective preference learning which is introduced in this paper. Furthermore, while the use of CNNs to process images is extensive, to the best of the authors knowledge, CNNs have not been applied before to process (or as a means to fuse) physiological signals.

As in many other machine learning applications, in affect detection it is common to apply dimensionality reduction techniques to the complete set of features extracted. A wide variety of feature selection (FS) methods have been used in the literature including sequential forward [31], sequential floating forward [10], sequential backwards [32], n-best individuals [33], perceptron [33] and genetic [34] feature selection. Fisher projection and Principal Component Analysis (PCA) have been also widely used as dimensionality reducers on different modalities of AC signals (e.g., see [10] among others). An auto-encoder can be viewed as a non-linear generalization of PCA [8]; however, while PCA has been applied in AC to transpose sets of manually extracted features into low dimensional spaces, in this paper auto-encoders are used to train unsupervised CNNs to transpose subsets of the raw input signals into a *learned* set of features. We expect that information relevant for prediction can be extracted more effectively using dimensionality reduction methods directly on the raw physiological signals than on a set of designer-selected extracted features.

B. Training Models of Affect

The selection of a method to create a model that maps a given set of features to predictions of affective variables is strongly influenced by the dynamic aspect of the features (stationary or sequential) and the format in which training examples are given (continuous values, class labels or ordinal labels). A vast set of *off-the-shelf* machine learning (ML) methods have been applied to create models of affect based on stationary features, irrespective of the specific emotions and modalities involved. These include Linear Discriminant Analysis [35], Multi-layer Perceptrons [32], K-Nearest Neighbors [36], Support Vector Machines [37], Decision Trees [38], Bayesian Networks [39], Gaussian Processes [29] and Fuzzy-rules [40]. On the other hand, Hidden Markov Models [41], Dynamic Bayesian Networks [42] and Recurrent Neural Networks [43] have been applied for constructing affect detectors that rely on features which change dynamically. In the approach presented here, deep neural network architectures reduce hierarchically the resolution of temporal signals down to a set of features that can be fed to simple stateless models eliminating the need for complex sequential predictors.

In all the above-mentioned studies, the prediction targets are either class labels or continuous values. Class labels are assigned either using an induction protocol (e.g., participants are asked to self-elicit an emotion [36], presented with stories to evoke a specific emotion [44]) or via rating- or rank-based questionnaires given to users experiencing the emotion (self-reports) or experts (third-person reports). If ratings are used, they can be binned into discrete or binary classes (e.g., on a scale from 1 to 5 measuring stress, values above or below 3 correspond to the user at stress or not at all, respectively [45]) or used as target values for supervised learning (e.g., two experts rate the amount of sadness of a facial expression and the average value is used as the sadness intensity [46]). Alternatively, if ranks are used, the problem of affective modeling becomes one of *preference learning*. In this paper we use *object ranking* methods—a subset of preference learning algorithms [47], [48]—which train computational models using partial orders among the training samples. These methods allow us to avoid binning together ordinal labels and to work with comparative questionnaires, which provide more reliable self-report data compared to ratings, as they generate less inconsistency and order effects [12].

Object ranking methods and comparative (rank) questionnaires have been scarcely explored in the AC literature, despite their well-known advantages. For example, Tognetti et al. [49] applied Linear Discriminant Analysis to learn models of preferences over game experiences based on physiological statistical features and comparative pairwise self-reports (i.e., participants played pairs of games and ranked games according to preference). On the same basis, Yannakakis et al. [50], [51] and Martínez et al. [34], [33] trained single and multiple layer perceptrons via genetic algorithms (i.e., *neuroevolutionary preference learning*) to learn models for several affective and cognitive states (e.g., fun, challenge and frustration) using physiological and behavioral data, and pairwise self-reports. In this paper we introduce a deep learning methodology for data given in a ranked

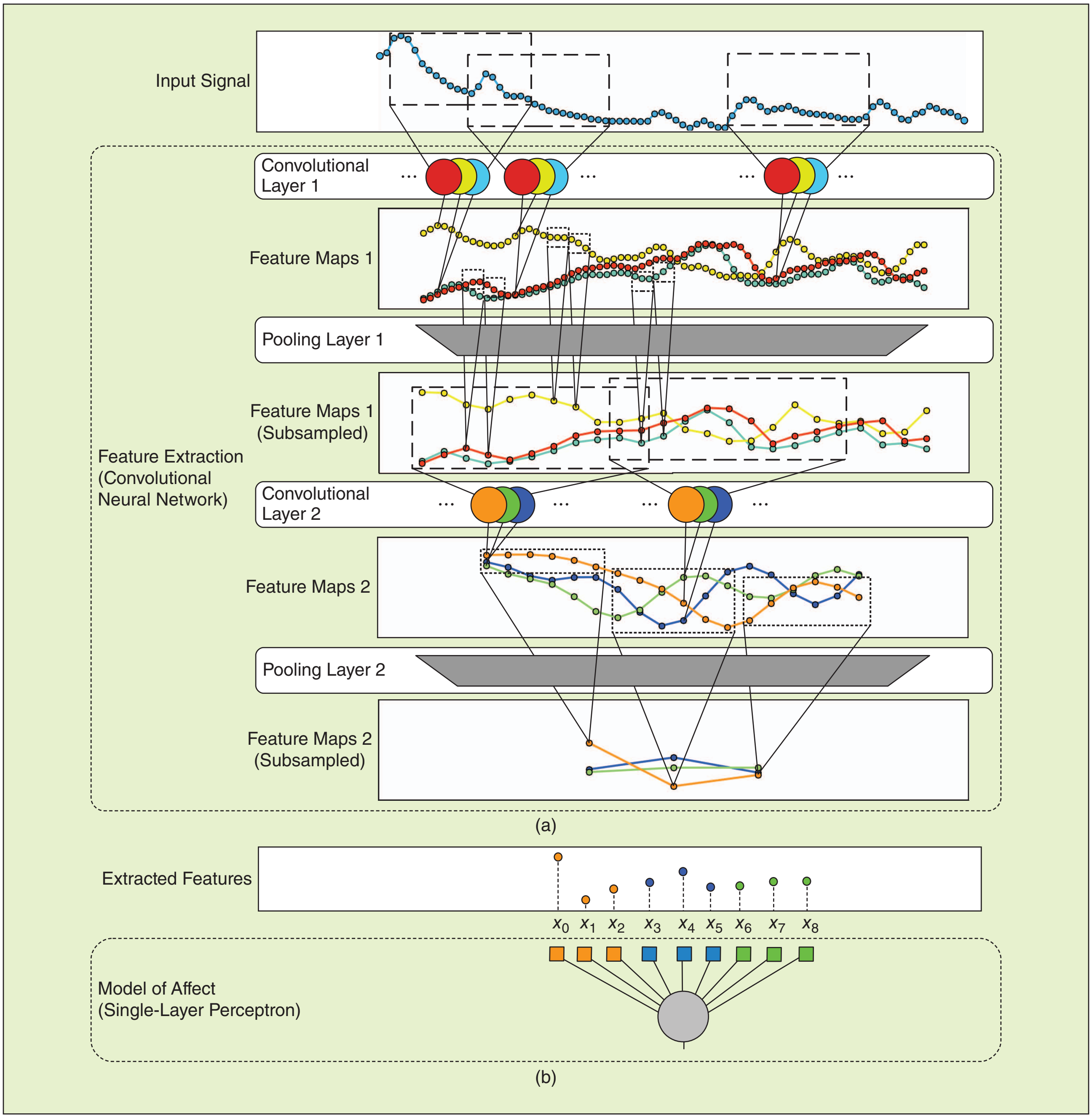

Input Signal
Convolutional Layer 1
Feature Maps 1
Pooling Layer 1
Feature Maps 1 (Subsampled)
Feature Extraction (Convolutional Neural Network)
Convolutional Layer 2
Feature Maps 2
Pooling Layer 2
Feature Maps 2 (Subsampled)
(a)
Extracted Features
x_0 x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8
Model of Affect (Single-Layer Perceptron)
(b)

FIGURE 1 Example of structure of a deep ANN architecture. The architecture contains: (a) a convolutional neural network (CNN) with two convolutional and two pooling layers, and (b) a single-layer perceptron (SLP) predictor. In the illustrated example the first convolutional layer (3 neurons and path length of 20 samples) processes a skin conductance signal which is propagated forward through an average-pooling layer (window length of 3 samples). A second convolutional layer (3 neurons and patch length of 11 samples) processes the subsampled feature maps and the resulting feature maps feed the second average-pooling layer (window length of 6 samples). The final subsampled feature maps form the output of the CNN which provides a number of extracted (learned) features which feed the input of the SLP predictor.

format (i.e., Preference Deep Learning) for the purpose of modeling affect.

III. Deep Artificial Neural Networks

We investigate an effective method of learning models that map signals of user behavior to predictions of affective states. To bypass the manual ad-hoc feature extraction stage, we use a *deep* model composed from (a) a multi-layer convolutional neural network (CNN) that transforms the raw signals into a reduced set of features that feed (b) a single-layer perceptron (SLP) which predicts affective states (see Fig. 1). Our hypothesis is that the automation of feature extraction via *deep learning* will yield physiological affect detectors of higher predictive power, which, in turn, will deliver affective models of higher accuracy. The advantages of deep learning techniques mentioned in the introduction of the paper have led to very promising results in computer vision as they have outperformed other state-of-the-art methods [52], [53]. Furthermore, convolutional networks have been successfully applied to dissimilar temporal datasets (e.g., [54], [25]) including electroencephalogram (EEG) signals [55] for seizure prediction.

To train the convolutional neural network (see Section III-A) we use *denoising auto-encoders* [56], an unsupervised learning method to train filters or feature extractors which transform the information of the input signal (see Section III-B) in order to capture a distributed representation of its leading factors of variation, but without the linearity assumption of PCA. The SLP is then trained using backpropagation [57] to map the outputs of the CNN to the given affective target values. In the case study examined in this paper, target values are given as pairwise comparisons (partial orders of length 2) making error functions commonly used with gradient descent methods, such as the difference of squared errors or cross-entropy, unsuitable for the task. For that purpose, we use the *rank margin* error function for preference data [58], [59] as detailed in Section III-C below. Additionally, we apply an automatic feature selection method to reduce the dimensionality of the feature space improving the prediction accuracy of the models trained (see Section III-D).

A. Convolutional Neural Networks

Convolutional or time-delay neural networks [25] are hierarchical models that alternate convolutional and pooling layers (see Fig. 1) in order to process large input spaces in which a spatial or temporal relation among the inputs exists (e.g., images, speech or physiological signals).

Convolutional layers contain a set of neurons that detect different patterns on a *patch* of the input (e.g., a time window in a time-series or part of an image). The inputs of each neuron (namely *receptive field*) determine the size of the patch. Each neuron contains a number of trainable weights equal to the number of its inputs and an additional *bias* parameter (also trainable); the output is calculated by applying an activation function (e.g., logistic sigmoid) to the weighted sum of the inputs plus the bias (see Fig. 2). Each neuron scans sequentially the input, assessing at each patch location the similarity to the pattern encoded on the weights. The consecutive outputs generated at every location of the input assemble a *feature map* (see Fig. 1). The output of the convolutional layer is the set of feature maps resulting from convolving each of the neurons across the input. Note that the convolution of each neuron produces the same number of outputs as the number of samples in the input signal (e.g., the sequence length) minus the size of the patch (i.e., the size of the receptive field of the neuron), plus 1 (see Fig. 1).

As soon as feature maps have been generated, a *pooling* layer aggregates consecutive values of the feature maps resulting from the previous convolutional layer, reducing their resolution with

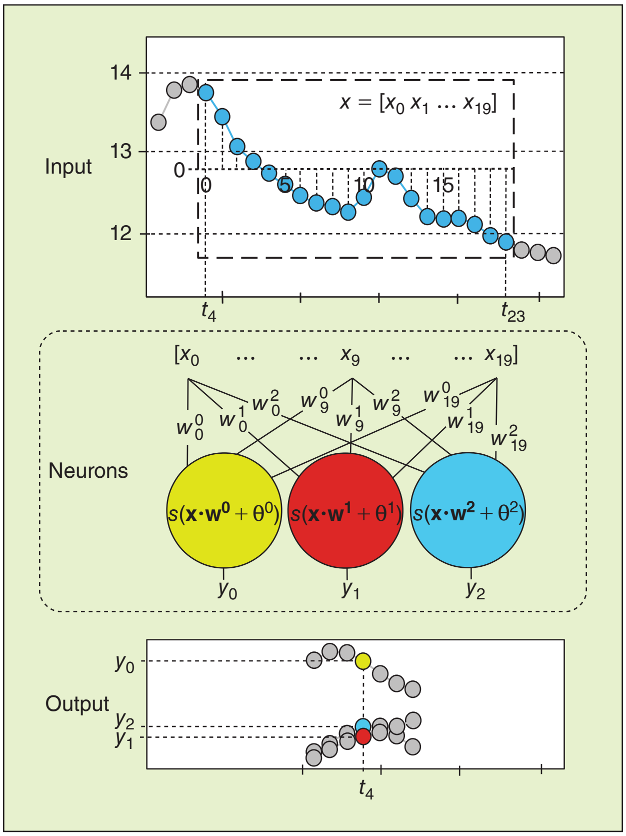

Input
14
13
12
0
0
5
10
15
$x = [x_0\ x_1 \dots x_{19}]$
t_4
t_{23}
$[x_0$ … … x_9 … … $x_{19}]$
w_0^0 w_0^1 w_0^2 w_9^0 w_9^1 w_9^2 w_{19}^0 w_{19}^1 w_{19}^2
Neurons
$s(\mathbf{x}\cdot\mathbf{w}^0+\theta^0)$
$s(\mathbf{x}\cdot\mathbf{w}^1+\theta^1)$
$s(\mathbf{x}\cdot\mathbf{w}^2+\theta^2)$
y_0 y_1 y_2
Output
y_0
y_2
y_1
t_4

FIGURE 2 Convolutional layer. The neurons in a convolutional layer take as input a patch on the input signal **x**. Each of the neurons calculates a weighted sum of the inputs (**x** · **w**), adds a bias parameter θ and applies an activation function *s(x)*. The output of each neuron contributes to a different feature map. In order to find patterns that are insensitive to the baseline level of the input signal, **x** is normalized with mean equal to 0. In this example, the convolutional layer contains 3 neurons with 20 inputs each.

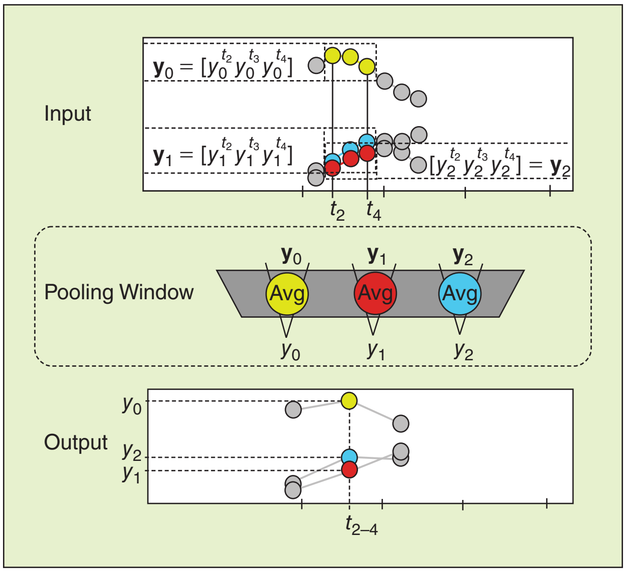

Input
$\mathbf{y}_0 = [y_0^{t_2}\ y_0^{t_3}\ y_0^{t_4}]$
$\mathbf{y}_1 = [y_1^{t_2}\ y_1^{t_3}\ y_1^{t_4}]$
$[y_2^{t_2}\ y_2^{t_3}\ y_2^{t_4}] = \mathbf{y}_2$
t_2 t_4
Pooling Window
$\mathbf{y}_0$ $\mathbf{y}_1$ $\mathbf{y}_2$
Avg Avg Avg
y_0 y_1 y_2
Output
y_0
y_2
y_1
$t_{2\text{–}4}$

FIGURE 3 Pooling layer. The input feature maps are subsampled independently using a pooling function over non-overlapping windows, resulting in the same number of feature maps with a lower temporal resolution. In this example, an average-pooling layer with a window length of 3 subsamples 3 feature maps.

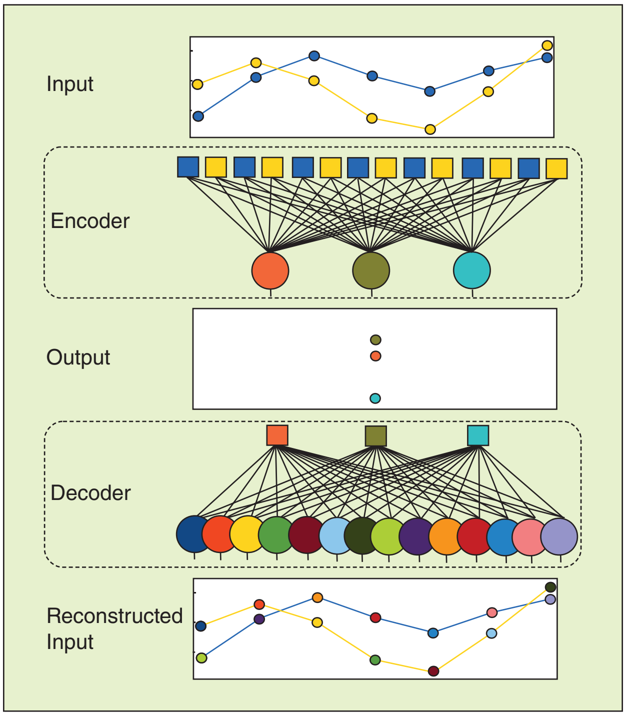


FIGURE 4 Structure of an auto-encoder. The encoder generates the learned representation (extracted features) from the input signals. During training the output representation is fed to a decoder that attempts to reconstruct the input.

a pooling function (see Fig. 3). The *maximum* or *average* values are the two most commonly used pooling functions providing *max-pooling* and *average-pooling* layers, respectively. This aggregation is typically done inside each feature map, so that the output of a pooling layer presents the same number of feature maps as its input but at a lower resolution (see Fig. 1).

B. Auto-Encoders

An auto-encoder (AE) [60], [8], [21] is a model that transforms an input space into a new distributed representation (extracted features) by applying a deterministic parametrized function (e.g., single layer of logistic neurons) called the encoder (see Fig. 4). The AE also learns how to map back the output of the encoder into the input space, with a parametrized decoder, so as to have small reconstruction error on the training examples, i.e., the original and corresponding decoded inputs are similar. However, constraints on the architecture or the form of the training criterion prevent the auto-encoder from simply learning the identity function everywhere. Instead, it will learn to have small reconstruction error on the training examples (and where it generalizes) and high reconstruction error elsewhere. Regularized auto-encoders are linked to density estimation in several ways [56], [61]; see [62] for a recent review of regularized auto-encoders. In this paper, the encoder weights (used to obtain the output representation) are also used to reconstruct the inputs (tied weights). By defining the reconstruction error as the sum of squared differences between the inputs and the reconstructed inputs, we can use a gradient descent method such as backpropagation to train the weights of the model. A denoising auto-encoder (DA) [56] is a variant of the basic model that during training adds a variable amount of noise to the inputs before computing the outputs. The resulting training objective is to reconstruct the original uncorrupted inputs, i.e., one minimizes the discrepancy between the outputs of the decoder and the original uncorrupted inputs.

Auto-encoders are among several unsupervised learning techniques that have provided remarkable improvements to gradient-descent supervised learning [4], especially when the number of labeled examples is small or in transfer settings [62]. ANNs that are *pretrained* using these techniques usually converge to more robust and accurate solutions than ANNs with randomly sampled initial weights. In this paper, we use a DA method known as Stacked Convolutional Auto-encoders [63] to train all convolutional layers of our CNNs from bottom to top. We trained the filters of each convolutional layer patchwise, i.e., by considering the input at each position (one patch) in the sequence as one example. This allows faster training than training convolutionally, but may yield translated versions of the same filter.

C. Preference Deep Learning

The outputs of a trained CNN define a number of *learned* features extracted from the input signal. These, in turn, may feed any function approximator or classifier that attempts to find a mapping between the input signal and a target output (i.e., affective state in our case). In this paper, we train a single layer perceptron to learn to predict the affective state of a user based on the learned features of her physiology (see Fig. 1). To this aim, we use backpropagation [57], which optimizes an error function iteratively across a number of epochs by adjusting the weights of the SLP proportionally to the gradient of the error with respect to the current value of the weights and current data samples.

We use the Rank Margin error function [64] that given two data samples $\{\mathbf{x_P},\mathbf{x_N}\}$ such that $\mathbf{x_P}$ is preferred over (or should be greater than) $\mathbf{x_N}$ is calculated as follows:

$$E(\mathbf{x_P},\mathbf{x_N}) = \max\{0, 1-(f(\mathbf{x_P})-f(\mathbf{x_N}))\}, \qquad (1)$$

where $f(\mathbf{x_P})$ and $f(\mathbf{x_N})$ represent the outputs of the SLP for the preferred and non-preferred sample, respectively. This function decreases linearly as the difference between the predicted value for preferred and non-preferred samples increases. The function becomes zero if this difference is greater than 1, i.e., there is enough margin to separate the preferred "positive example" score $f(\mathbf{x_P})$ from the nonpreferred "negative example" score $f(\mathbf{x_N})$. By minimizing this function, the neural network is driven towards learning outputs separated at least by one unit of distance between the preferred and non preferred data sample. In each training epoch, for every pairwise preference in the training dataset, the output of the neural network is computed for the two data samples in the preference (preferred and non preferred) and the rank-margin error is backpropagated through the network in order to obtain the gradient required to update the weights. Note that while all layers of the deep architecture could be trained (including supervised fine-tuning of the CNNs), due to the small number of labeled examples available here, the Preference Deep Learning algorithm is constrained to the last layer (i.e., SLP) of the network in order to avoid over fitting.

D. Automatic Feature Selection

Automatic feature selection (FS) is an essential process towards picking those features (deep learned or ad-hoc extracted) that are appropriate for predicting the examined affective states. In this paper, we use Sequential Forward Feature Selection (SFS) for its low computational effort and demonstrated good performance compared to more advanced, nevertheless time consuming, feature subset selection algorithms such as the genetic-based FS [34]. While a number of other FS algorithms are available for comparison, in this paper we focus on the comparative benefits of learned physiological detectors over ad-hoc designed features. The impact of FS on model performance is further discussed in Section VI.

In brief, SFS is a bottom-up search procedure where one feature is added at a time to the current feature set (see e.g., [48]). The feature to be added is selected from the subset of the remaining features such that the new feature set generates the maximum value of the performance function over all candidate features for addition. Since we are interested in the minimal feature subset that yields the highest performance, we terminate selection procedure when an added feature yields equal or lower validation performance to the performance obtained without it. The performance of a feature set selected by automatic FS is measured through the average classification accuracy of the model in three independent runs using 3-fold cross-validation. In the experiments presented in this paper, the SFS algorithm selects the input feature set for the SLP model.

IV. The Maze-Ball Dataset

The dataset used to evaluate the proposed methodology was gathered during an experimental game survey where 36 participants played four pairs of different variants of the same video-game. The test-bed game named *Maze-Ball* is a 3D prey/predator game that features a ball inside a maze controlled by the arrow keys. The goal of the player is to maximize her score in 90 seconds by collecting a number of pellets scattered in the maze while avoiding enemies that wander around. Eight different game variants were presented to the players. The games were different with respect to the virtual camera profile used, which determined how the virtual world was presented on screen. We expected that different camera profiles would induce different experiences and affective states, which would, in turn, reflect on the physiological state of the players, making it possible to predict the players' affective self-reported preferences using information extracted from their physiology.

Blood volume pulse (BVP) and skin conductance (SC) were recorded at 31.25 Hz during each game session. The players filled in a 4-alternative forced choice questionnaire after completing a pair of game variants reporting whether the first or the second game of the pair (i.e., pairwise preference) felt more *anxious, exciting, frustrating, fun* and *relaxing,* with options that include equally or none at all [33]. While three additional labels were collected in the original experiment (*boredom, challenge* and *frustration*), we focus only on affective states or states that are implicitly linked to affective experiences, such as fun (thereby, removing the cognitive state of *challenge*), and report only results for states in which prediction accuracies of over 70% were achieved in at least one of the input feature sets examined (thereby, removing *frustration*). Finally, *boredom* was removed due to the small number of clear preferences available (i.e., most participants reported not feeling bored during any of the games). The details of the Maze-Ball game design and the experimental protocol followed can be found in [33], [34].

A. Ad-Hoc Extraction of Statistical Features

This section lists the statistical features extracted from the two physiological signals monitored. Some features are extracted for both signals while some are signal-dependent as seen in the list below. The choice of those specific statistical features is made in order to cover a fair amount of possible BVP and SC signal dynamics (tonic and phasic) proposed in the majority of previous studies in the field of psychophysiology (e.g., see [15], [65], [51] among many).

- ❑ **Both signals** ($\alpha \in \{\mathrm{BV\,P}, \mathrm{SC}\}$): Average $E\{\alpha\}$, standard deviation $\sigma\{\alpha\}$, maximum $\max\{\alpha\}$, minimum $\min\{\alpha\}$, the difference between maximum and minimum signal recording $D^{\alpha} = \max\{\alpha\} - \min\{\alpha\}$, time when maximum α occurred $t_{\max}\{\alpha\}$, time when minimum α occurred $t_{\min}\{\alpha\}$ and the difference $D_t^{\alpha} = t_{\max}\{\alpha\} - t_{\min}\{\alpha\}$; autocorrelation (lag equals 1) of the signal ρ_1^{α} and mean of the absolute values of the first and second differences of the signal [15] $\delta_{|1|}^{\alpha}$ and $\delta_{|2|}^{\alpha}$ respectively).
- ❑ **BVP:** Average inter-beat amplitude $E\{\mathrm{IBAmp}\}$; given the inter-beat time intervals (RR intervals) of the signal, the following Heart Rate Variability (HRV) parameters were computed: the standard deviation of RR intervals $\sigma\{\mathrm{RR}\}$, the fraction of RR intervals that differ by more than 50 msec from the previous RR interval $p\mathrm{RR}50$ and the root-mean-square of successive differences of RR intervals $\mathrm{RMS}_{\mathrm{RR}}$ [65].
- ❑ **SC:** Initial, $\mathrm{SC}_{\mathrm{in}}$, and last, $\mathrm{SC}_{\mathrm{last}}$, SC recording, the difference between initial and final SC recording $D_{l-i}^{\mathrm{SC}} = \mathrm{SC}_{\mathrm{last}} - \mathrm{SC}_{\mathrm{in}}$ and Pearson's correlation coefficient R_{SC} between raw SC recordings and the time t at which data were recorded.

V. Experiments

To test the efficacy of DL on constructing accurate models of affect we pretrained several convolutional neural networks—using denoising auto-encoders—to extract features for each of the physiological signals and across all reported affective states in the dataset. The topologies of the networks were selected after preliminary experiments with 1- and 2-layer CNNs and

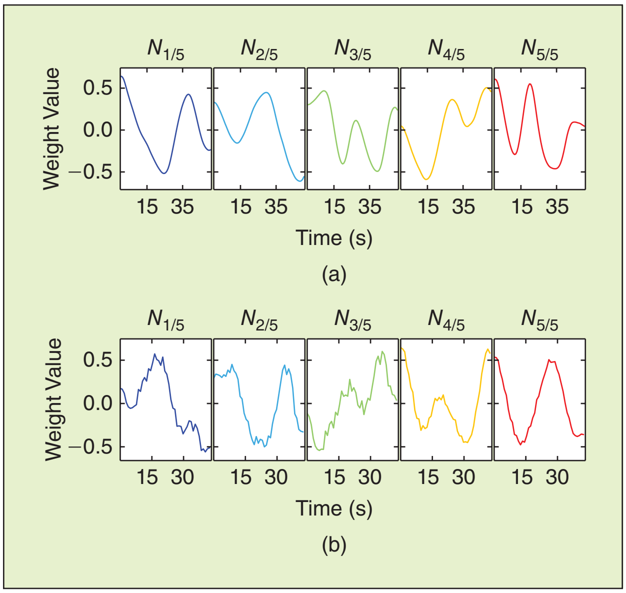


FIGURE 5 Learned features of the best-performing convolutional neural networks. Lines are plotted connecting the values of consecutive connection weights for each neuron N_x. The x axis displays the time stamp (in seconds) of the samples connected to each weight within the input patch. (a) $\mathrm{CNN}^{\mathrm{SC}}_{80}$ (skin conductance). (b) $\mathrm{CNN}^{\mathrm{BVP}}_{1\times45}$ (blood volume pulse).

trained using the complete unlabeled dataset. In all experiments reported in this paper the final number of features pooled from the CNNs is 15, to match the number of ad-hoc extracted statistical features (see Section IV-A). Although a larger number of pooled features could potentially yield higher prediction accuracies, we restricted the size to 15 to ensure a fair comparison against the accuracies yielded by the ad-hoc extracted features.

The input signals are not normalized using global, baseline or subject-dependent constants; instead, the first convolutional layer of every CNN subtracts the mean value within each patch presented, resulting in patches with a zero mean value inside the patch, making learned features that are only sensitive to variation within the desired time window (patch) and insensitive to the baseline level (see Fig. 2). As for statistical features, we apply z-transformation to the complete dataset: the mean and the standard deviation value of each feature in the dataset are 0 and 1, respectively. Independently of model input, the use of preference learning models—which are trained and evaluated using within-participant differences—automatically minimizes the effects of between-participants physiological differences (as noted in [33], [12] among other studies).

We present a comparison between the prediction accuracy of several SLPs trained either on the learned features of the CNNs or on the ad-hoc designed statistical features. The affective models are trained with and without automatic feature selection and compared. This section presents the key findings derived from the SC (Section V-A) and the BVP (Section V-B) signals and concludes with the analysis of the fusion of the two physiological signals (Section V-C). All the experiments presented here run for 10 times and the average (and standard error) of the resulting models' prediction accuracies are reported. The prediction accuracy of the models is calculated as the average 3-fold cross-validation (CV) accuracy (average percentage of correctly classified pairs on each fold). While more folds in cross-validation (e.g., 10) or other validation methods such as leave-one-out cross-validation are possible, we considered the 3-fold CV as appropriate for testing the generalizability of the trained ANNs given the relatively small size of (and the high across-subject variation existent in) this dataset.

A. Skin Conductance

The focus of the paper is on the effectiveness of DL for affective modeling. While the topology of the CNNs can be critical for the performance of the model, the exhaustive empirical validation of all possible CNN topologies and parameter sets is out of the scope of this paper. For this purpose—and also due to space considerations—we have systematically tested critical parameters of CNNs (e.g., the patch length, the number of layers, and the number of neurons), we have fixed a number of CNN parameters (e.g., pooling window length) based on suggestions from the literature and we discuss results from representative CNN architectures. In particular, for the skin conductance signal we present results on two pretrained CNNs. The first, labeled $\mathrm{CNN}^{\mathrm{SC}}_{20\times11}$, contains two convolutional layers with 5 logistic neurons per patch location at each layer, as well as average-pooling over non-overlapping windows of size 3. Each of the neurons in the first and second convolutional layer has 20 and 11 inputs, respectively. The second network (labeled as $\mathrm{CNN}^{\mathrm{SC}}_{80}$), contains one convolutional layer with 5 logistic neurons of 80 inputs each, at each patch location.

Both CNNs examined here are selected based on a number of criteria. The number of inputs of the first convolutional layer of the two CNNs considered were selected to extract features at different time resolutions (20 and 80 inputs corresponding to 12.8 and 51.2 seconds, respectively) and, thereby, giving an indication of the impact the time resolution might have on performance. Extensive experiments with smaller and larger time windows did not seem to affect the model's prediction accuracy. The small window on the intermediate pooling layer was chosen to minimize the amount of information lost from the feature maps while the number of inputs to the neurons in the next layer was adjusted to cover about a third of the pooled feature maps. Finally, we selected 5 neurons in the first convolutional layer as a good compromise between expressivity and dissimilarity among the features learned: a low number of neurons derived features with low expressivity while a large number of neurons generally resulted in features being very similar.

Both topologies are built on top of an average-pooling layer with a window length of 20 samples and are topped up with an average-pooling layer that pools 3 outputs per neuron. Although SC is usually sampled at high frequencies (e.g., 256 Hz), we believe that the most affect-relevant information contained in the signal can be found at a lower time resolutions as even rapid arousal changes (i.e., a phasic change of SC) can be captured with a lower resolution and at a lower computational cost [66], [33]. For that purpose, the selection of this initial pooling stage aims to facilitate feature learning at a resolution of 1.56 Hz. Moreover, experiments with dissimilar pooling layers showed that features extracted on higher SC resolutions do not necessarily yield models of higher accuracy. The selection of 5 neurons for the last convolutional layer and the following pooling layer was made to achieve the exact number of ad-hoc statistical features of SC (i.e. ,15).

1) Deep Learned Features

Figure 5(a) depicts the values of the 80 connection weights of the five neurons in the convolutional layer of the $\mathrm{CNN}^{\mathrm{SC}}_{80}$ which cover 51.2 seconds of the SC signal (0.64 seconds per weight) on each evaluation. The first neuron (N_1) outputs a maximal value for areas of the SC signal in which a long decay is followed by 10 seconds of an incremental trend and a final decay. The second neuron (N_2) shows a similar pattern but the increment is detected earlier in the time window and the follow-up decay is longer. A high output of these neurons would suggest that a change in the experience elicited a heightened level of arousal that decayed naturally seconds after. The forth neuron (N_4) in contrast, detects a second incremental trend in the signal that elevates the *SC* level even further. The fifth neuron (N_5) also detects two increments but several seconds further apart. Finally, the third neuron (N_3) detects three consecutive *SC* increments. These last three neurons could detect changes on the level of arousal caused by consecutive stimuli presented few seconds apart. Overall, this convolutional layer captures long and slow changes (10 seconds or more) of skin conductance. These local patterns cannot be modeled with the same precision using standard statistical features related to variation (such as standard deviation and average first/second absolute differences), which further suggests that dissimilar aspects of the signal are extracted by learned and ad-hoc features.

2) DL vs. Ad-Hoc Feature Extraction

Figure 6(a) depicts the average prediction accuracies (3-fold CV) of SLPs trained on the outputs of the CNNs compared to the corresponding accuracies obtained by SLPs trained on the ad-hoc extracted statistical features. Both CNN topologies yield predictors of relaxation with accuracies over 60% (66.07% and 65.38% for $\mathrm{CNN}^{\mathrm{SC}}_{20\times11}$, and $\mathrm{CNN}^{\mathrm{SC}}_{80}$, respectively), which are significantly higher than the models built on statistical features. Given the performance differences among these networks, it appears that learned local features could detect aspects of SC that were more relevant to the prediction of this particular affective state than the set of ad-hoc statistical features proposed. Models trained on automatically selected features further validate this result [see Fig. 6(b)] showing differences with respect to statistical features above 5%. Furthermore, the relaxation models trained on selected ad-hoc features, despite the benefits of FS, yield accuracies lower than the models trained on the complete sets of learned features. This suggests that CNNs can extract general information from SC that is more relevant for affect modeling than statistical features selected specifically for the task. An alternative interpretation is that the feature space created by CNNs allows backpropagation to find more general solutions than the greedy-reduced (via SFS) space of ad-hoc features.

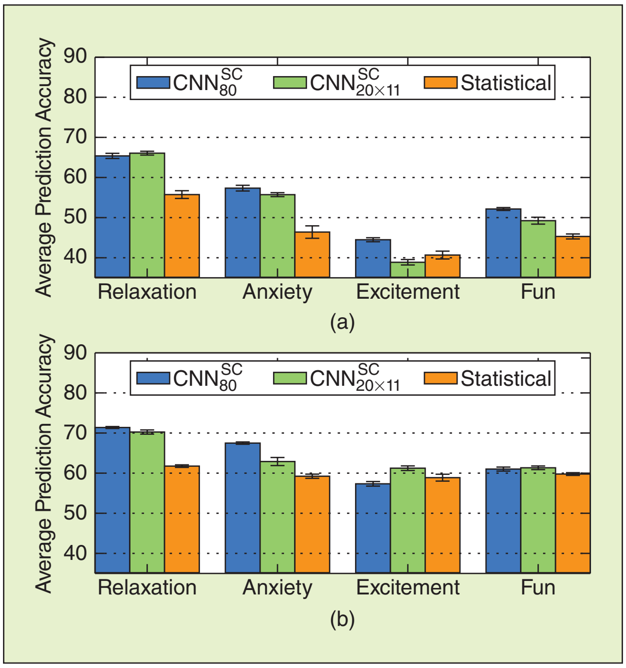


FIGURE 6 Skin conductance: average accuracy of SLPs trained on statistical features (statistical), and features pooled from each of the CNN topologies ($\mathrm{CNN}^{\mathrm{SC}}_{20\times11}$ and $\mathrm{CNN}^{\mathrm{SC}}_{80}$). The black bar displayed on each average value represents the standard error (10 runs). (a) All features. (b) Features selected via SFS.

For all other emotions considered, neither the CNNs nor the ad-hoc statistical features lead to models that can significantly improve chance prediction (see [67] for random baselines on this dataset). When feature selection is used [see Fig. 6(b)], CNN-based models outperform statistical-based models on the prediction of every affective state with accuracies above 60% with at least one topology.

Despite the difficulty of predicting complex affective states based solely on SC, these results suggest that unsupervised CNNs trained as a stack of denoising auto-encoders form a promising method to automatically extract features from this modality, as higher prediction accuracies were achieved when compared against a well-defined set of ad-hoc statistical features. Results also show that there are particular affective states (*relaxation* and *anxiety*, to a lesser degree), in which DL is able to automatically extract features that are beneficial for their prediction. On the other hand, it appears that DL has a lesser effect in predicting some affective states (*fun* and *excitement*) based on the SC signal compared to models build on the ad-hoc designed features. Prediction accuracies in those affective states for both type

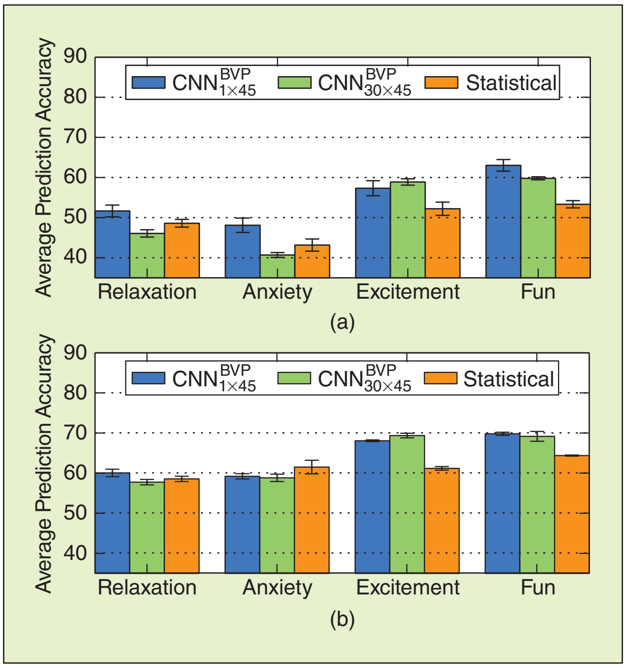


FIGURE 7 Blood volume pulse: average accuracy of SLPs trained on statistical features (statistical), and features pooled from each of the CNN topologies ($\text{CNN}^{\text{BVP}}_{1\times 45}$ and $\text{CNN}^{\text{BVP}}_{30\times 45}$). The black bar displayed on each average value represents the standard error (10 runs). (a) All features. (b) Features selected via SFS.

of features (ad-hoc or CNN-extracted) are rather low, suggesting that SC is not an appropriate signal for their modeling in this dataset. It is worth mentioning that earlier studies on this dataset [67] report higher accuracies on the ad-hoc statistical features than those reported here. In that study, however, two different signal components were extracted from the SC signal, leading to three times the number of features examined in this paper (i.e., 45 features). Given the results obtained in this paper, it is anticipated that by using more learned features—for example, combining CNNs with different input lengths that would capture information from different time resolutions—DL can reach and surpass those baseline accuracies.

B. Blood Volume Pulse

Following the same systematic approach for selecting CNN topology and parameter sets, we present two convolutional networks for the experiments on the Blood Volume Pulse (BVP) signal. The CNN architectures used in the experiments feature the following: 1) one max-pooling layer with nonoverlapping windows of length 30 followed by a convolutional layer with 5 logistic neurons per patch location and 45 inputs at each neuron ($\text{CNN}^{\text{BVP}}_{1\times 45}$); and 2) two convolutional layers with 10 and 5 logistic neurons per patch location, respectively, and an intermediate max-pooling layer with a window of length 30. The neurons of each layer contain 30 and 45 inputs, respectively ($\text{CNN}^{\text{BVP}}_{1\times 45}$). As in the CNNs used in the SC experiments, both topologies are topped up with an average-pooling layer that reduces the length of the outputs from each of the 5 output neurons down to 3—i.e., the CNNs output 5 feature maps of length 3 which amounts to 15 features. The initial pooling layer of the first network collects the maximum value of the BVP signal every 0.96 seconds, which results in an approximation of the signal's *upper envelope*—that is a smooth line joining the extremes of the signal's peaks. Decrements in this function are directly linked with increments in heart rate (HR), and further connected with increased arousal and corresponding affective states (e.g., excitement and fun [33], [18]). Neurons with 45 inputs were selected to capture long patterns (i.e., 43.2 seconds) of variation, as sudden and rapid changes in heart rate were not expected during the experiment game survey. The second network follows the same rationale but the first pooling layer—instead of collecting the maximum of the raw BVP signal—processes the outputs of 10 neurons that analyze signal patches of 0.96 seconds, which could operate as a beat detector mechanism.

1) Deep Learned Features

Figure 5(b) depicts the 45 connection weights of each neuron in $\text{CNN}^{\text{BVP}}_{1\times 45}$ which cover 43.2 seconds of the BVP signal's upper envelope. Given the negative correlation between the trend of the BVP's upper envelope and heart rate, neurons produce output of maximal values when consecutive decreasing weight values are aligned with a time window containing an HR increment and consecutive increasing weight values with HR decays. On that basis, the second (N_2) and fifth (N_5) neurons detect two 10-second-long periods of HR increments, which are separated by an HR decay period. The first (N_1) and the forth (N_4) neuron detect two overlapping increments on HR, followed by a decay in N_4. The third neuron (N_3), on the other hand, detects a negative trend on HR with a small peak in the middle. This convolutional layer appears to capture dissimilar local complex patterns of BVP variation which are, arguably, not available through common ad-hoc statistical features.

2) DL vs. Ad-Hoc Feature Extraction

Predictors of excitement and fun trained on features extracted with $\text{CNN}^{\text{BVP}}_{1\times 45}$ outperformed the ad-hoc feature sets—both the complete [see Fig. 7(a)] and the automatically selected feature sets [see Fig. 7(b)]. It is worth noting that no other model improved baseline accuracy using all features [see Fig. 7(a)]. In particular, excitement and fun models based on statistical features achieved performances of 61.1% and 64.3%, respectively, which are significantly lower than the corresponding accuracies of $\text{CNN}^{\text{BVP}}_{1\times 45}$ [68.0% and 69.7 %, respectively—see Fig. 7(b)] and not significantly different from the accuracies of $\text{CNN}^{\text{BVP}}_{1\times 45}$ with the complete set of features [57.3% and 63.0%, respectively—see Fig. 7(a)]. Given the reported links between fun and heart rate [18], this result suggests that $\text{CNN}^{\text{BVP}}_{1\times 45}$ effectively extracted HR information from the BVP signal to predict reported fun. The efficacy of CNNs is further supported by the results reported in [67] where SLP predictors of fun trained on statistical features of the HR signal (in the same dataset examined here) do not outperform the DL models presented in this paper. For reported fun and excitement, CNN-based feature extraction demonstrates a great advantage of extracting affect-relevant information from BVP bypassing beat detection and heart rate estimation.

Models built on selected features for relaxation and anxiety yielded low accuracies around 60%, showing small differences between learned and ad-hoc features, which suggests that BVP-based emotional manifestations are not the most appropriate predictors for those two states in this dataset. Despite the challenges that the periodicity of blood volume pulse generates in affective modeling, CNNs managed to extract powerful features to predict two affective states, outperforming the statistical features proposed in the literature and matching more complex data processing methods used in similar studies [67].

C. Fusion of SC and BVP

To test the effectiveness of learned features in fused models, we combined the outputs of the BVP and SC CNN networks presented earlier into one SLP and compared its performance against a combination of all ad-hoc BVP and SC features. For space considerations we only present the combination of the best performing CNNs trained on each signal individually—i.e., CNN^{SC}_{80} and $\text{CNN}^{BVP}_{1\times45}$. The fusion of CNNs from both signals generates models that yield higher prediction accuracies than models built on ad-hoc features across all affective states, using both all features and subsets of selected features (see Fig. 8). This result further validates the effectiveness of CNNs for modeling affect from physiological signals, as models trained on automatically selected learned features from the two signals yield prediction accuracies around 70-75%. In all cases but one (i.e., anxiety prediction with SFS) these performances are significantly higher than the performances of corresponding models built on commonly used ad-hoc statistical features.

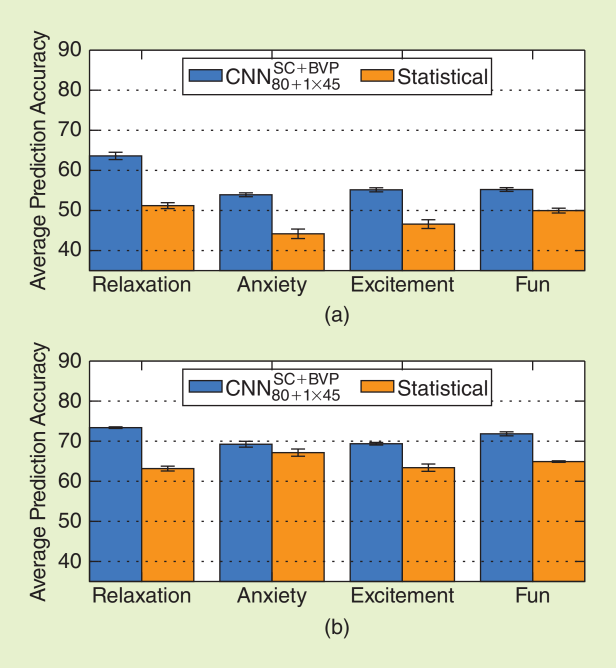


FIGURE 8 Fusion of SC and BVP signals: average accuracy of SLPs trained on blood volume pulse and skin conductance using statistical features on the raw signal (statistical) and features pooled from CNN^{SC}_{80} and $\text{CNN}^{BVP}_{1\times45}$ $\text{CNN}^{SC+BVP}_{80+1\times45}$. The black bar displayed on each average value represents the standard error (10 runs). (a) All features. (b) Features selected via SFS.

VI. Discussion

Even though the results obtained are more than encouraging with respect to the applicability and efficacy of DL for affective modeling, there are a number of research directions that should be considered in future research. While the Maze-Ball game dataset includes key components for affective modeling and is representative of a typical affective modeling scenario, our PDL approach needs to be tested on diverse datasets. The reduced size of the dataset limited the number of features that could be learned. Currently, deep architectures are widely used to extract thousands of features from large datasets, which yields models that outperform other state-of-the-art classification or regression methods (e.g., [27]). We expect that the application of DL to model affect in large physiological datasets would show larger improvements with respect to statistical features and provide new insights on the relationship between physiology and affect. Moreover, to be able to demonstrate robustness of the algorithm, more and dissimilar modalities of user input need to be considered, and different domains (beyond games) need to be explored. To that end, different approaches to multimodal fusion in conjunction with DL need to be investigated. The accuracies obtained across different affective states and modalities of user input, however, already provide sufficient evidence that the method would generalize well in dissimilar domains and modalities.

The paper did not provide a thorough analysis of the impact of feature selection to the efficiency of DL as the focus was put on feature extraction. To that end, more feature selection methods will need to be investigated and compared to SFS. While ad-hoc feature performances might be improved with more advanced FS methods, such as genetic-search based FS [34], the obtained results already show that DL matches and even beats a rather effective and popular FS mechanism without the use of feature selection in several experiments. Although in this paper we have compared DL to a complete and representative set of ad-hoc features, a wider set of features could be explored in future work. For instance, heart rate variability features derived from the Fourier transformation of BVP (see [33]) could be included in the comparison. However, it is expected that CNNs would be able to extract relevant frequency-based features as their successful application in other domains already demonstrates (e.g., music sample classification [54]). Furthermore, other automatic feature extraction methods, such as principal component analysis, which is common in domains, such as image classification [68], will be explored for psycho-physiological modeling and compared to DL in this domain.

Learned features derived from DL architectures may define data-based extracted patterns, which could lead to the advancement of our understanding of emotion manifestations via physiology.

Despite the good results reported in this paper on the skin conductance and blood volume pulse signals, we expect that certain well-designed ad-hoc features can still outperform automatically learned features. Within playing behavioral attributes, for example, the final score of a game—which is highly correlated to reported fun in games [69]—may not be captured by convolutional networks, which tend to find patterns that are invariant with respect to the position in the signal. Such an ad-hoc feature, however, may carry information of high predictive power for particular affective states. We argue that DL is expected to be of limited use in low resolution signals (e.g., player score over time) which could generate well-defined feature spaces for affective modeling.

An advantage of ad-hoc extracted statistical features resides in the simplicity to interpret the physical properties of the signal as they are usually based on simple statistical metrics. Therefore, prediction models trained on statistical features can be analyzed with low effort providing insights in affective phenomena. Artificial neural networks have traditionally been considered as *black boxes* that oppose their high prediction power to a more difficult interpretation of what has been learned by the model. We have shown, however, that appropriate visualization tools can ease the interpretation of neural-network based features. Moreover, learned features derived from DL architectures may define data-based extracted patterns, which could lead to the advancement of our understanding of emotion manifestations via physiology (and beyond).

Finally, while DL can automatically provide a more complete and appropriate set of features when compared to adhoc feature extraction, parameter tuning is a necessary phase in (and a limitation of) the training process. This paper introduced a number of CNN topologies that performed well on the SC and BVP signals while empirical results showed that, in general, the performance of the CNN topologies is not affected significantly by parameter tuning. Future work, however, would aim to further test the sensitivity of CNN topologies and parameter sets as well as the generality of the extracted features across physiological datasets, reducing the experimentation effort required for future applications of DL to psychophysiology.

VII. Conclusions

This paper introduced the application of deep learning (DL) to the construction of reliable models of affect built on physiological manifestations of emotion. The algorithm proposed employs a number of convolutional layers that learn to extract relevant features from the input signals. The algorithm was tested on two physiological signals (skin conductance and blood volume pulse) individually and on their fusion for predicting the reported affective states of *relaxation, anxiety, excitement* and *fun* (given as pairwise preferences). The dataset is derived from 36 players of a 3D prey/predator game. The proposed preference deep learning (PDL) approach overcomes standard ad-hoc feature extraction used in the affective computing literature as it manages to yield models of equal or significantly higher prediction accuracy across all affective states examined. The increase in performance is more evident when automatic feature selection is employed.

Results, in general, suggest that DL methodologies are highly appropriate for affective modeling and, more importantly, indicate that ad-hoc feature extraction can be redundant for physiology-based modeling. Furthermore, in some affective states examined (e.g., relaxation models built on SC; fun and excitement models built on BVP; relaxation models built on fused SC and BVP), DL without feature selection manages to reach or even outperform the performances of models built on ad-hoc extracted features which are boosted by automatic feature selection. These findings showcased the potential of DL for affective modeling, as both manual feature extraction and automatic feature selection could be ultimately bypassed.

With small modifications, the methodology proposed can be applied for affect classification and regression tasks across any type of input signal. Thus, the method is directly applicable for affect detection in one-dimensional time-series input signals such as electroencephalograph (EEG), electromyograph (EMG) and speech, but also in two-dimensional input signals such as images [27] (e.g., for facial expression and head pose analysis). Finally, results suggest that the method is powerful when fusing different type of input signals and, thus, it is expected to perform equally well across multiple modalities.

Acknowledgment

The authors would like to thank Tobias Mahlmann for his work on the development and administration of the cluster used to run the experiments. Special thanks for proofreading goes to Yana Knight. Thanks also go to the Theano development team, to all participants in our experiments, and to Ubisoft, NSERC and Canada Research Chairs for funding. This work is funded, in part, by the ILearnRW (project no: 318803) and the C2Learn (project no. 318480) FP7 ICT EU projects.

References

[1] R. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 2000.

[2] R. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affective Comput.*, vol. 1, no. 1, pp. 18–37, 2010.

[3] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[4] Y. Bengio, "Learning deep architectures for AI," *Found. Trends® Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[5] I. Arel, D. Rose, and T. Karnowski, "Deep machine learning–A new frontier in artificial intelligence research [Research Frontier]," *IEEE Comput. Intell. Mag.*, vol. 5, no. 4, pp. 13–18, Nov. 2010.

[6] M. Dash and H. Liu, "Feature selection for classification," *Intell. data anal.*, vol. 1, nos. 1-4, pp. 131–156, 1997.

[7] M. Ranzato, F. Huang, Y. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2007, pp. 1–8.

[8] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[9] Y. Bengio and O. Delalleau, "On the expressive power of deep architectures," in *Algorithmic Learning Theory*. Berlin, Germany: Springer-Verlag, 2011, pp. 18–36.

[10] E. Vyzas and R. Picard, "Affective pattern classification," in *Proc. AAAI 1998 Fall Symp. Emotional Intelligent: The Tangled Knot Cognition*, pp. 176–182, 1998.

[11] H. P. Martínez and G. N. Yannakakis, "Mining multimodal sequential patterns: A case study on affect detection," in *Proc. 13th. Int. Conf. Multimodal Interfaces*, 2011, pp. 3–10.

[12] G. N. Yannakakis and J. Hallam, "Ranking vs. preference: A comparative study of self-reporting," in *Proc. 4th Int. Conf. Affective Computing Intelligent Interaction,* 2011, pp. 437–446.

[13] J. Morris, "Observations: SAM: The self-assessment manikinan efficient cross-cultural measurement of emotional response," *J. Advertising Res.*, vol. 35, no. 6, pp. 63–68, 1995.

[14] J. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, p. 1161, 1980.

[15] R. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp. 1175–1191, 2001.

[16] D. Ververidis and C. Kotropoulos, "Automatic speech classification to five emotional states based on gender information," in *Proc. Eusipco*, Vienna, pp. 341–344, 2004.

[17] D. Giakoumis, D. Tzovaras, K. Moustakas, and G. Hassapis, "Automatic recognition of boredom in video games using novel biosignal moment-based features," *IEEE Trans. Affective Comput.*, vol. 2, no. 3, pp. 119–133, July-Sept. 2011.

[18] G. N. Yannakakis and J. Hallam, "Entertainment modeling through physiology in physical play," *Int. J. Human-Comput. Stud.*, vol. 66, no. 10, pp. 741–755, Oct. 2008.

[19] S. Pincus, "Approximate entropy as a measure of system complexity," in *Proc. National Academy Sciences*, 1991, vol. 88, no. 6, pp. 2297–2301.

[20] N. Lesh, M. Zaki, and M. Ogihara, "Mining features for sequence classification," in *Proc. 5th ACM Int. Conf. Knowledge Discovery Data Mining*, 1999, pp. 342–346.

[21] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layerwise training of deep networks," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press 2007, vol. 19, p. 153.

[22] P. Ekman and W. Friesen, "Facial action coding system: A technique for the measurement of facial movement," in *From Appraisal to Emotion: Differences Among Unpleasant Feelings, Motivation and Emotion*, P. C. Ellsworth, and C. A. Smith, Eds. Palo Alto, CA: Consulting Psychologists Press, 1988, vol. 12, pp. 271–302.

[23] G. Caridakis, S. Asteriadis, K. Karpouzis, and S. Kollias, "Detecting human behavior emotional cues in natural interaction," in *Proc. 17th Int. Conf. Digital Signal Processing*, July 2011, pp. 1–6.

[24] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Trans. Affective Comput.*, vol. PP, no. 99, p. 1, 2012.

[25] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA: MIT Press 1995, vol. 3361, pp. 255–258.

[26] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural Netw.*, vol. 16, no. 5, pp. 555–559, 2003.

[27] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *Proc. European Conf. Computer Vision*, 2012, pp. 802–822.

[28] J. Susskind, A. Anderson, and G. E. Hinton, "The Toronto face dataset," U. Toronto, Toronto, ON, Canada, Tech. Rep. UTML TR 2010-001, 2010.

[29] A. Kapoor, W. Burleson, and R. Picard, "Automatic prediction of frustration," *Int. J. Human-Comput. Stud.*, vol. 65, no. 8, pp. 724–736, 2007.

[30] S. Tognetti, M. Garbarino, A. Bonarini, and M. Matteucci, "Modeling enjoyment preference from physiological responses in a car racing game," in *Proc. IEEE Conf. Computational Intelligence Games*, 2010, pp. 321–328.

[31] C. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Processing,* vol. 13, no. 2, pp. 293–303, 2005.

[32] J. Wagner, J. Kim, and E. André, "From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2005, pp. 940–943.

[33] G. N. Yannakakis, H. P. Martínez, and A. Jhala, "Towards affective camera control in games," *User Model. User-Adapted Interact.*, vol. 20, no. 4, pp. 313–340, 2010.

[34] H. P. Martínez and G. N. Yannakakis, "Genetic search feature selection for affective modeling: A case study on reported preferences," in *Proc. 3rd Int. Workshop Affective Interaction Natural Environments*, 2010, pp. 15–20.

[35] D. Giakoumis, A. Drosou, P. Cipresso, D. Tzovaras, G. Hassapis, T. Zalla, A. Gaggioli, and G. Riva, "Using activity-related behavioural features towards more effective automatic stress detection," *PLoS ONE*, vol. 7, no. 9, p. e43571, 2012.

[36] O. AlZoubi, R. Calvo, and R. Stevens, "Classification of EEG for affect recognition: An adaptive approach," in *AI 2009 Proc. 22nd Australasian Joint Conf. Advances in Artificial Intelligence,* pp. 52–61. 2009.

[37] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Trans. Affective Comput.*, vol. 3, no. 2, pp. 211–223, 2012.

[38] S. Mcquiggan, B. Mott, and J. Lester, "Modeling self-efficacy in intelligent tutoring systems: An inductive approach," *User Model. User-Adapted Interact.*, vol. 18, no. 1, pp. 81–123, 2008.

[39] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *J. Netw. Comput. Appl.*, vol. 30, no. 4, pp. 1334–1345, 2007.

[40] R. Mandryk and M. Atkins, "A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies," *Int. J. Human-Comput. Stud.*, vol. 65, no. 4, pp. 329–347, 2007.

[41] J. F. Grafsgaard, K. E. Boyer, and J. C. Lester, "Predicting facial indicators of confusion with hidden Markov models," in *Affective Computing and Intelligent Interaction*, (Series Lecture Notes in Computer Science), S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Berlin, Germany: Springer-Verlag, 2011, vol. 6974, pp. 97–106.

[42] R. Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Real-Time Vision Human-Computer Interaction*. New York: Springer-Verlag, 2005, pp. 181–200.

[43] H. Kobayashi and F. Hara, "Dynamic recognition of basic facial expressions by discrete-time recurrent neural network," in *Proc. Int. Joint Conf. Neural Networks*, Oct. 1993, vol. 1, pp. 155–158.

[44] K. Kim, S. Bang, and S. Kim, "Emotion recognition system using short-term monitoring of physiological signals," *Med. Biol. Eng. Comput.*, vol. 42, no. 3, pp. 419–427, 2004.

[45] J. Hernandez, R. R. Morris, and R. W. Picard, "Call center stress recognition with person-specific models," in *Affective Computing and Intelligent Interaction*, (Series Lecture Notes in Computer Science), S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Berlin, Germany: Springer-Verlag, 2011, vol. 6974, pp. 125–134.

[46] J. Bailenson, E. Pontikakis, I. Mauss, J. Gross, M. Jabon, C. Hutcherson, C. Nass, and O. John, "Real-time classification of evoked emotions using facial feature tracking and physiological responses," *Int. J. Human-Computer Stud.*, vol. 66, no. 5, pp. 303–317, 2008.

[47] J. Fürnkranz and E. Hüllermeier, "Preference learning," *Künstliche Intell.*, vol. 19, no. 1, pp. 60–61, 2005.

[48] G. N. Yannakakis, "Preference learning for affective modeling," in *Proc. Int. Conf. Affective Computing Intelligent Interaction*, Amsterdam, The Netherlands, Sept. 2009, pp. 126–131.

[49] S. Tognetti, M. Garbarino, A. Bonanno, M. Matteucci, and A. Bonarini,"Enjoyment recognition from physiological data in a car racing game," in *Proc. 3rd Int. Workshop Affective Interaction Natural Environments*, 2010, pp. 3–8.

[50] G. N. Yannakakis, J. Hallam, and H. H. Lund, "Entertainment capture through heart rate activity in physical interactive playgrounds," *User Model. User-Adapted Interact.*, vol. 18, no. 1, pp. 207–243, 2008.

[51] G. N. Yannakakis and J. Hallam, "Entertainment modeling through physiology in physical play," *Int. J. Human-Comput. Stud.*, vol. 66, no. 10, pp. 741–755, 2008.

[52] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*. Cambridge, MA: MIT Press, 2012.

[53] C. Farabet, C. Couprie, L. Najman, Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, p. 1–15, 2013.

[54] P. Hamel, S. Lemieux, Y. Bengio, and D. Eck, "Temporal pooling and multiscale learning for automatic annotation and ranking of music audio," in *Proc. 12th Int. Conf. Music Information Retrieval*, 2011, pp. 729–734.

[55] P. Mirowski, Y. LeCun, D. Madhavan, and R. Kuzniecky, "Comparing SVM and convolutional networks for epileptic seizure prediction from intracranial EEG," in *Proc. IEEE Workshop Machine Learning Signal Processing*, 2008, pp. 244–249.

[56] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. Int. Conf. Machine Learning*, 2008, pp. 1096–1103.

[57] D. Rumelhart, *Backpropagation: Theory, Architectures, and Applications*. Hillsdale, NJ: Lawrence Erlbaum, 1995.

[58] B. Bai, J. Weston, D. Grangier, R. Collobert, K. Sadamasa, Y. Qi, O. Chapelle, and K. Weinberger, "Learning to rank with (a lot of) word features," *Inform. Retrieval*, vol. 13, no. 3, pp. 291–314, 2010.

[59] D. Grangier and S. Bengio, "Inferring document similarity from hyperlinks," in *Proc. ACM Int. Conf. Information Knowledge Management*, 2005, pp. 359–360.

[60] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length, and Helmholtz free energy," in *Proc. Neural Information Processing System NIPS'1993*, 1994, pp. 3–10.

[61] G. Alain, Y. Bengio, and S. Rifai, "Regularized auto-encoders estimate local statistics," Dept. IRO, Université de Montréal, Montreal, QC, Canada, Tech. Rep. Arxiv Report 1211.4246, 2012.

[62] Y. Bengio, A. Courville, and P. Vincent, "Unsupervised feature learning and deep learning: A review and new perspectives," Université de Montréal, Tech. Rep. Arxiv Report 1206.5538, 2012.

[63] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Proc. Int. Conf. Artificial Neural Networks and Machine Learning*, pp. 52–59, 2011.

[64] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. Int. Conf. Machine Learning*, 2008, pp. 160–167.

[65] J. Goldberger, S. Challapalli, R. Tung, M. Parker, and A. Kadish, "Relationship of heart rate variability to parasympathetic effect," *Circulation*, vol. 103, no. 15, p. 1977, 2001.

[66] N. Ravaja, T. Saari, M. Salminen, J. Laarni, and K. Kallinen, "Phasic emotional reactions to video game events: A psychophysiological investigation," *Media Psychol.*, vol. 8, no. 4, pp. 343–367, 2006.

[67] H. P. Martínez, M. Garbarino, and G. N. Yannakakis, "Generic physiological features as predictors of player experience," *Affective Comput. Intell. Interact.*, pp. 267–276, 2011.

[68] W. Zhao, R. Chellappa, and A. Krishnaswamy, "Discriminant analysis of principal components for face recognition," in *Proc 3rd IEEE Int. Conf. IEEE Automatic Face Gesture Recognition*, 1998, pp. 336–341.

[69] H. P. Martínez, K. Hullett, and G. N. Yannakakis, "Extending neuro-evolution preference learning through player modeling," in *Proc. IEEE Conf. Computational Intelligence and Games*, Copenhagen, Denmark, Aug. 2010, pp. 313–320.