IEEE TRANSACTIONS ON CYBERNETICS, VOL. 52, NO. 5, MAY 2022 3097

Multimodal Vigilance Estimation Using Deep Learning

Wei Wu, Member, IEEE, Wei Sun, Q. M. Jonathan Wu, Senior Member, IEEE, Yimin Yang, Senior Member, IEEE, Hui Zhang, Member, IEEE, Wei-Long Zheng, Member, IEEE, and Bao-Liang Lu, Senior Member, IEEE

Abstract—The phenomenon of increasing accidents caused by reduced vigilance does exist. In the future, the high accuracy of vigilance estimation will play a significant role in public transportation safety. We propose a multimodal regression network that consists of multichannel deep autoencoders with subnetwork neurons (MCDAEs). After we define two thresholds of “0.35” and “0.70” from the percentage of eye closure, the output values are in the continuous range of 0–0.35, 0.36–0.70, and 0.71–1 representing the awake state, the tired state, and the drowsy state, respectively. To verify the efficiency of our strategy, we first applied the proposed approach to a single modality. Then, for the multimodality, since the complementary information between forehead electrooculography and electroencephalography features, we found the performance of the proposed approach using features fusion significantly improved, demonstrating the effectiveness and efficiency of our method.

Index Terms—Deep learning, dimension reduction, electroencephalography (EEG), electrooculography (EOG), multimodal vigilance estimation.

I. INTRODUCTION

DIFFERENT groups of scientists use the term “vigilance” in different ways [1]. In psychiatry, vigilance can be specifically described as attention to a potentially dangerous condition, with hypervigilance that is a symptom of anxiety disorder [2]. In the area of cognitive neuroscience, vigilance refers to the ability to focus on a task within a lasting time [3]. In clinical neurophysiology, the vigilance level is related to sleep–wake [4]. In the above discussion, the most common definition is that vigilance represents sustained attention.

The Centers for Disease Control and Prevention (CDC) [5] reported that 4.2% of the 147,076 adult respondents from nearly 20 states or regions reported having had at least one drowsy drive in the past 30 days in 2009–2010. Reduced or complete loss of vigilance has been resulting in an increasing number of traffic accidents around the world [6], [7], meaning it should be taken seriously. Various methods of studying vigilance, including subjective methods [8], behavioral methods [9]–[11], and vehicle-based methods [12], have been proposed to cope with this problem. However, the primary limitations of those methods arise from ignoring the uniqueness of the individual driver and neglecting the personal biases involved, as well as the monotony of the simulated environment under experimental conditions.

Physiological signals or nonvisual features of drivers with healthy physical conditions have fewer false than visual features and can be used to predict drowsiness in a timely manner. In fact, the methods based on physiological signals that represent internal cognitive states have been gradually considered as an efficient means of assessing vigilance. Scientists have found a certain relationship between electrocardiography (ECG) and fatigue, including heart rate (HR) decrease and HR variability
(HRV), changes during fatigue [13], [14], and a healthy subject with prolonged fatigue has the reduced respiration rate (RR) value [15]. The study of electromyography (EMG) shows that the frequency spectrum shifts to low frequencies and the amplitude of the EMG signal increases when muscles become fatigued [16], [17]. Electroencephalography (EEG) [18], [19] is generated from the potentials that are recorded from the rhythmic activity of the postsynaptic cortical neuron, which is synchronized by the complex interaction of a large number of cortical cells. Among various physiological indicators, EEG is considered to be the most important and reliable because it directly records the neurophysiological signal of the human brain.

In addition to EEG, electrooculography (EOG) evaluates various eye movements, which can provide valuable warning indications of drowsiness. EOG is another promising measure of assessing vigilance [20]. Unlike the traditional EOG (EOGt)-based method, a simple but more robust method proposed by Zheng and Lu [21] utilizes the placement of wearable electrodes on the forehead area. The amplitude of forehead EOG (EOGf) is significantly lower after extraction by a median filter, and suitable electrodes placement reduces the user’s discomfort.

Huo et al. [22] and Wu et al. [23] have proposed the multimodal methods for estimating the level of vigilance and achieving better performance. For example, Huo et al. [22] used a fusion strategy that employs feature-level fusion (FLF) to detect fatigue levels, combined with a graph-regularized extreme learning machine (GELM). The average value of the correlation coefficient (COR)/root mean-square error (RMSE) greatly improved, moving to 0.81/0.07 using fusion signals, while the corresponding average COR/RMSE values were 0.73/0.09 for single modality.

We propose using a network of multichannel deep autoencoders with subnetwork neurons (MCDAEm) to obtain the optimal features, employing feature fusion to estimate vigilance. Here, we use four different EOG and EEG datasets from SEED-VIG. Our work contributes to the research on the topic as follows.

1) Compared to the other existing iterative deep-learning (DL)-based networks [24], instead of being randomly acquired, the hidden layers of the proposed autoencoder model are calculated by replacement technologies and include only four steps. Simultaneously, the proposed architecture aims for dimension reduction and signal reconstruction instead of relying on efficient classification applications, as do the other existing multilayer network models.

2) Unlike the other traditional multilayer networks [25], the proposed model consists of many hidden nodes, each of which can be considered as a layer of the network model and has capabilities of feature selection and representation learning. Simultaneously, the input data are randomly divided into five batches, each of which through processes of dimension reduction, subspace feature extraction, and subspace feature combination.

3) To quantify vigilance, the output values are a series of continuous value in the range of 0 to 1 corresponding to the percentage of eye closure (PERCLOS). Blink components, such as impulses from vertical EOG (EOGv) feature and saccade components from horizontal EOG (EOGh) feature that can be easily detected by the proposed algorithm, which is consistent with our previous conclusions [21].

II. RELATED WORKS

A. Description of the Dataset

Fig. 1(a) shows the data-collection apparatus, wherein the experimental vehicle is an engineless car in which the gas pedal and steering wheel are controlled by software. An LCD screen in front of the vehicle simulates a highway driving environment and is updated in real time. Subjects signed written informed consent before participating, and this research was approved by the local ethics committee. The data were gathered from 23 human subjects, including 11 men and 12 women with an average age of 23. All of the subjects were healthy, with normal hearing, visual acuity that was normal.
or corrected to normal, no visible trauma to the head, no use of medication, and no addiction to alcohol or tobacco, and they were all given regular rest according to the timetable. To ensure that the subject was in a drowsy state while driving, the average time for the experiment was approximately 2 h, from 12:30 P.M. to 2:30 P.M., with the human subjects ing, the average time for the experiment was approximately to ensure that the subject was in a drowsy state while driv-
ing indicator of arousal states, which has been widely explored
arousal is not causal. Instead, eye movement acts as a promis-
ing feedback occurred. Fig. 1(b) displays the SensoMotoric
subject fell asleep during the experiment, no real-time warn-
ing vector, and no addiction to alcohol or tobacco, and
wider recognition in sleep-deprived persons. In public driving safety fields, the
PERCLOS is one of the most widely acknowledged vigilance indicators in [10] and [32]–[34]. The PERCLOS algorithm, introduced by Wierwille [32], has a mean that is the propor-
tion of time for which eyes remain closed in a given unit of
time. Grace and Davis [34] at Carnegie Mellon University repeated this finding, using a high-resolution cam-
era for testing an eye closure over a specific value to judge drowsiness. Fig. 2 illustrates the
PERCLOS based on the eyes closed degree (ECD), and PERCLOS can be calculated as follows:

\[ V_P = \frac{T_2 - T_1}{T_3 - T_0}, \quad V_P \in [0, 1] \] (2)

where \( V_P \) represents the value of PERCLOS; and \((T_2 - T_1)\) and \((T_3 - T_0)\) indicate the eyes closure duration and the duration time from 20% closed state to 20% open state, respectively.

We found that the method mentioned above only pays atten-
tion to two states—eyes closed and eyes open—instead of all
the important eye movements that provide crucial information
for estimating vigilance. Simultaneously, the performance of
the method based on traditional facial videos [35] can easily
be influenced by environmental factors, especially brightness
and occlusion. Thus, we use an automatic continuous vigilance
annotation method [36] that employs SMI-ETGW, offering up
to 120-Hz high resolution. SMI-ETGW can more fully reflect
eye movements, including blinks, saccades, and fixation com-
ponents, and the PERCLOS training labels calculated by it
can be regarded as an accurate and feasible ocular parame-
ter for real-time testing fatigue. This approach can be used
for dual tasks in both laboratory and real-world environments. The formula is as follows:

\[ V_P = \frac{T_2 - T_1}{T_3 - T_0}, \quad V_P \in (0, 1) \]

\[ T_{duration} = \frac{T_2 - T_1}{T_b + T_s + T_f} \] (3)

where \( T_b, T_s, \) and \( T_f \) represent the time of blink, saccade, and
fixation state, respectively.

2) Forehead EOG: The eye movements we have analyzed
in this study were spontaneous, rather than intentional. The

B. Previous Related Network Models

As an important algorithm in artificial intelligence (AI), neu-
nal network-based algorithms have been widely used for EEG
signal processing. Our two extended neural network models
are proposed for the vigilance estimation, such as deep autoen-
coder (DAE) [20] and multilayer autoencoder (MAE) [27]. We used EOG-based single-modal DAE to estimate vigilance and achieve an accuracy of 80%. Meanwhile, Yang et al. proposed MAE for image reconstruction and dimension reduc-
tion with the Moore–Penroze inverse matrix learning strategy. The features could be compressed by a single-hidden layer
with extremely fast processing speed.

Autoencoders can only address single-type samples and
can cause beneficial representations of the inputs; however,
we argue that a better representation should also depend on
the internal relationship between the input pairs. Thus, we
proposed MCDAE \( s \) that can handle multiple types of samples.
The formula of output is

\[ \mathbf{H} = S(a_i, b_j, x_i), \quad i = \{1, 2, \ldots, d\}, \quad j = \{1, 2, \ldots, n\} \] (1)

where \( j \) and \( i \) represent the \( j \)th subnetwork nodes and its \( i \)th
hidden nodes, respectively.

Each subnetwork node is only connected to its adjacent,
which can be considered as an independent system, improv-
ing learning efficiency effectively. In addition, the subnetwork
neurons as subspace feature extractors significantly increase
the generalization performance [28], as long as the generated
subspace features could be mixed and combined in the late
stages for classification.

III. METHODOLOGY

A. Data Preprocessing

1) PERCLOS: In continuous vigilance assessment with a
supervised machine learning paradigm, the chief challenge is
how to quantitatively mark physiological signals that are col-
lected from the sensors because it is theoretically difficult to
accurately obtain the ground truth of the transformed phys-
iological state. The association between eye movements and
arousal is not causal. Instead, eye movement acts as a promis-
ing indicator of arousal states, which has been widely explored
in previous studies. For example, in the neuroscience field,
Wang et al. [29] proposed that spontaneous eyelid closures
can serve as a proxy for vigilance and be jointly analyzed
with functional magnetic resonance imaging (fMRI) [30], [31]
to determine vigilance fluctuation. Spontaneous eyelid clo-
sure also served as a good marker of reduced responsiveness
in sleep-deprived persons. In public driving safety fields, the
PERCLOS is one of the most widely acknowledged vigilance
indicators in [10] and [32]–[34]. The PERCLOS algorithm,
introduced by Wierwille [32], has a mean that is the propor-
tion of time for which eyes remain closed in a given unit
of time. Grace and Davis [34] at Carnegie Mellon University
repeatedly verified this finding, using a high-resolution cam-
era for testing an eye closure over a specific value to judge
drowsiness. Fig. 2 illustrates the measuring principle of
the PERCLOS based on the eyes closed degree (ECD), and
PERCLOS can be calculated as follows:

\[ V_P = \frac{T_2 - T_1}{T_3 - T_0}, \quad V_P \in [0, 1] \] (2)

where \( V_P \) represents the value of PERCLOS; and \((T_2 - T_1)\) and
\((T_3 - T_0)\) indicate the eyes closure duration and the duration
time from 20% closed state to 20% open state, respectively.

We found that the method mentioned above only pays atten-
tion to two states—eyes closed and eyes open—instead of all
the important eye movements that provide crucial information
for estimating vigilance. Simultaneously, the performance of
the method based on traditional facial videos [35] can easily
be influenced by environmental factors, especially brightness
and occlusion. Thus, we use an automatic continuous vigilance
annotation method [36] that employs SMI-ETGW, offering up
to 120-Hz high resolution. SMI-ETGW can more fully reflect
eye movements, including blinks, saccades, and fixation com-
ponents, and the PERCLOS training labels calculated by it
can be regarded as an accurate and feasible ocular parame-
ter for real-time testing fatigue. This approach can be used
for dual tasks in both laboratory and real-world environments. The formula is as follows:

\[ V_P = \frac{T_2 - T_1}{T_3 - T_0}, \quad V_P \in (0, 1) \]

\[ T_{duration} = \frac{T_2 - T_1}{T_b + T_s + T_f} \] (3)

where \( T_b, T_s, \) and \( T_f \) represent the time of blink, saccade, and
fixation state, respectively.

2) Forehead EOG: The eye movements we have analyzed
in this study were spontaneous, rather than intentional. The

Fig. 2. Details of ECD.
Fig. 3. EOG and EEG data collected from different electrode placements. (a) Forehead electrode positions. (b) Temporal EEG and posterior EEG were recorded from 6-channel electrode positions and 12-channel electrode positions, respectively.

**TABLE I**

<table>
<thead>
<tr>
<th>Components</th>
<th>Sequence</th>
<th>Encode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative peak</td>
<td>Single</td>
<td>0</td>
</tr>
<tr>
<td>Positive peak</td>
<td>Single</td>
<td>1</td>
</tr>
<tr>
<td>Blink candidate</td>
<td>Three successive</td>
<td>010</td>
</tr>
<tr>
<td>Saccade candidate</td>
<td>Two successive</td>
<td>01 or 10</td>
</tr>
</tbody>
</table>

**TABLE II**

<table>
<thead>
<tr>
<th>EOG</th>
<th>EOG&lt;sub&gt;horizontal&lt;/sub&gt; by ICA</th>
<th>EOG&lt;sub&gt;horizontal&lt;/sub&gt; by MIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>EOG&lt;sub&gt;vertical&lt;/sub&gt; by ICA</td>
<td>EOG&lt;sub&gt;vertical&lt;/sub&gt;-ICA</td>
<td>EOG&lt;sub&gt;vertical&lt;/sub&gt;-MIN &amp; EOG&lt;sub&gt;vertical&lt;/sub&gt;-ICA</td>
</tr>
<tr>
<td>EOG&lt;sub&gt;vertical&lt;/sub&gt; by MIN</td>
<td>EOG&lt;sub&gt;vertical&lt;/sub&gt;-ICA &amp; EOG&lt;sub&gt;vertical&lt;/sub&gt;-MIN</td>
<td>EOG&lt;sub&gt;vertical&lt;/sub&gt;-MIN</td>
</tr>
</tbody>
</table>

**TABLE III**

<table>
<thead>
<tr>
<th>Features</th>
<th>Dimension</th>
<th>Sample</th>
<th>Data format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saccade</td>
<td>13</td>
<td>885</td>
<td>36 x 885</td>
</tr>
<tr>
<td>Fixation</td>
<td>10</td>
<td>885</td>
<td></td>
</tr>
<tr>
<td>Blink</td>
<td>13</td>
<td>885</td>
<td></td>
</tr>
</tbody>
</table>
intentionally controlled by subjects, which causes degraded performance in prediction. We therefore combined different modalities in our study by including brain signals. EEG is an electrical activity that is noninvasively recorded from the scalp, which cannot be intentionally controlled by subjects. Moreover, the changes in brain activities contribute to early warnings of reduced vigilance. Herein lies the motivation for our development of a multimodal machine learning algorithm combining both EOG and EEG to estimate vigilance. According to the traditional methods [41], [42], ocular artifacts (OAs), including eye movements and blinks, are always considered the most dominant type of contamination—especially for the signals that are collected from the frontal head, which produces higher—magnitude signals, allowing them to travel throughout the scalp, distorting and masking EEG signals. Compared with conventional approaches to removing EOG through the Gaussian distribution, and

\[
f(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2}.
\]

(6)

Then, we obtained the DE features \( h(x | \mu, \sigma^2) \) as follows:

\[
h(x | \mu, \sigma^2) = -\int_{-\infty}^{\infty} f(x | \mu, \sigma^2) \log f(x | \mu, \sigma^2) \, dx
\]

\[
= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2} \log \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2} \, dx
\]

\[
= \frac{1}{2} \log \left(\frac{\sqrt{2\pi} \sigma}{2}\right) e^{\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2}.
\]

(7)

\[
T_{\text{EEG}} \quad \text{Moving average (MA)} \quad \text{Linear dynamic system (LDS)}
\]

<table>
<thead>
<tr>
<th>DE</th>
<th>DE-MA</th>
<th>DE-LDS</th>
<th>PSD</th>
<th>PSD-MA</th>
<th>PSD-LDS</th>
</tr>
</thead>
</table>

Fig. 4. Double-layer network in the proposed framework. The \( m \)-dimensional features \( F_e \) are obtained by mapping \( n \)-dimensional input data \( X \).

its probability density function. The formula can be expressed as follows:

\[
f(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2}.
\]

(6)

Then, we obtained the DE features \( h(x | \mu, \sigma^2) \) as follows:

\[
h(x | \mu, \sigma^2) = -\int_{-\infty}^{\infty} f(x | \mu, \sigma^2) \log f(x | \mu, \sigma^2) \, dx
\]

\[
= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2} \log \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2} \, dx
\]

\[
= \frac{1}{2} \log \left(\frac{\sqrt{2\pi} \sigma}{2}\right) e^{\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2}.
\]

(7)

In addition, we extracted DE, and power spectral density (PSD) features are extracted by the total frequency bands with 2-Hz frequency resolution (2 Hz) between 1 and 50 Hz and five frequency bands (5Bands), respectively. After using the moving average (MA) and the linear dynamic system (LDS) filtering, we listed the EEG features used, and their formats are in Tables IV and V, respectively. For example, in Table IV, PSD-MA represented PSD features extracted by raw data prior to the use of the MA separation strategy.

B. Network Model

According to the advantage of physiological signals, we know that the EOG \( f \) has two properties: 1) a high ratio of signal to noise and 2) ease of setting up. Meanwhile, EEG can fully record neurophysiological signals about vigilance.
EOG and EEG signals have complementary characteristics. We proposed a network MCDAE to test the accuracy of vigilance. The related notations are defined in Table VI.

1) Subspace Feature Dimension Reduction and Signal Reconstruction: A double-layer network in the proposed framework is described in Fig. 4. The low-dimensional features $F_e$ are obtained by mapping high-dimensional input data $X$. The weights of the decoder layer are used as the input data of the next encoding layer and all parameters used by the encoding layer are updated (see Fig. 5). The raw input data were divided into five batches, each through the processes of dimension reduction and signal reconstruction. The details can be described as follows:

Step 1: Given $N$ arbitrary distinct training samples from a continuous system $\{([X_n, \ Y_n])_i | X \in \mathbb{R}^n, Y \in (0, 1)\}$. Output data were similar to the input data, which were reconstructed by an autoencoder. Here $X = Y$, we then used the activation function $\phi_{1}(x) = \sin(x)$ in the encoding layer. The initial general input weights ($W_{ej}, W_{ej} \in \mathbb{R}^{n_1 \times m_1}$) and biases ($b_{ej}, b_{ej} \in \mathbb{R}$) were orthogonal random generation, as follows:

$$F_{e1} = \sum_{i=1}^{d} F_{ei1} = \phi_{1}(W_{e11} \cdot X, b_{e11})$$

$$= \sum_{i=1}^{d} \sin(W_{i1} \cdot X + b_{e1})$$

$$W_{ej} \cdot W_{ej} = I, \ b_{ej} \cdot b_{ej} = 1$$

(8)

where $F_{e1}$ is current feature data.

Step 2: The inverse of the activation function is $\phi_{1}^{-1}(x) = \arcsin(x)$. The optimal parameters for the $j$th decoding layer $\{(W_{dj}, b_{dj}) | W_{dj} \in \mathbb{R}^{n_1 \times m_1}, b_{dj} \in \mathbb{R}\}$ are obtained by

$$W_{dj1} = \phi_{1}^{-1}(Y) \cdot F_{e1j} = \arcsin(Y) \cdot F_{e1j}$$

$$b_{dj1} = \text{RMSE}(W_{dj1} \cdot F_{e1} - \arcsin(Y))$$

(9)

where RMSE represents the root-mean-squared error that is the standard deviation (SD) of the residuals.

According to the ridge regression (RR) technique, the diagonal elements of the matrix $FF'$ or $F'F$ should contain the shrinkage ridge parameter ($K > 0$) in a multiple regression analysis. The inverse function of Moore–Penrose $F'$ can be expressed as follows:

$$F' = \begin{cases} (FF')^{-1}F & \text{if } (FF') \text{ is nonsingular} \\ (K/I + FF')^{-1}F' & \text{if } (FF') \text{ is singular} \end{cases}$$

(10)

Step 3: Set $j = j + 1$, update $W_{ej}, b_{ej}$, and $F_{ej}$ by

$$W_{ej} = W_{dj}$$

$$b_{ej} = b_{dj}$$

$$F_{ej} = \phi_{1}(W_{ej} \cdot X + b_{ej})$$

(11)

Step 4: We repeat steps 2 and 3 ($n - 1$) times, obtaining the parameters of $W_{ej}, W_{ej}, W_{ej},$ and the feature $F_e$.

2) Subspace Feature Extraction: The initial feature of the $j$th (the initial index $j = 1$) subnetwork neurons in a subspace feature extraction layer is obtained from step 4. We found that the initial feature is $F_{ej} = F_e$.

Step 5: Given the $\phi_{2}(x) = 1/(1 + \exp^{-x})$ and $\phi_{3}(x) = 1/(1 + \exp^{-x})$ activation functions of the entrance layer and

---

**TABLE VI**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>$k = e, \ k = d, \ k = 2, \text{ and } k = 3$ represent encoding layer, decoding layer, subspace feature extraction layer and subspace feature combination layer, respectively.</td>
</tr>
<tr>
<td>$W_{ij}$</td>
<td>$W_{ij}$ is the $i^{th}$ hidden neuron's feature of the $j^{th}$ encoding layer $(k = e)$ or decoding layer $(k = d)$; is the $j^{th}$ subnetwork neuron $(k = 2 \text{ or } k = 3)$ in $k^{th}$ layer. $(W_{kl}, \ W_{kl} \in \mathbb{R}^{d \times m})$ and $(b_{kl}, \ b_{kl} \in \mathbb{R})$ represent its weight and bias, respectively.</td>
</tr>
<tr>
<td>$F_{k}$</td>
<td>$F_{k}$ is the Moore–Penrose inverse of the matrix $F$.</td>
</tr>
<tr>
<td>$m_k$</td>
<td>$m_k$ represents input data dimension of $k^{th}$ layer.</td>
</tr>
<tr>
<td>$\phi_{e}(x)$</td>
<td>$\phi_{e}(x) = \sin$, $\phi_{e}(x) = \text{sigmoid}$, and $\phi_{e}(x) = \text{sigmoid}$ represent activation functions of the dimension reduction layer, subspace feature extraction layer and subspace feature combination layer, respectively.</td>
</tr>
</tbody>
</table>
exit layer, respectively, we obtained the features of the jth subnetwork neurons \( \{W_{3j}, b_{3j}\}, W_3 \in \mathbb{R}^{m_1 \times m_2}, b_3 \in \mathbb{R} \) by

\[
W_{3j} = \phi_3^{-1}(L(Y)) \cdot F_{2j}
\]

\[
= -(\log(1/L(Y) - 1))/(1 + \exp(-(W_{2j}, x_{b2})^T)
\]

\[
b_{3j} = \text{RMSE} \left( W_{2j} \cdot (1 + \exp(-(W_{2j}, x_{b2}))) + \log(1/L(Y) - 1) \right). 
\]

(12)

Step 6: We updated \( e_j, W_{2j}, \) and \( b_{2j} \) as follows:

\[
e_j = Y - L^{-1} \cdot \phi_3(1 + \exp(-(W_{2j}, x_{b2}))
\]

\[
W_{2j} = \phi_3^{-1}(L(P_{j-1} + F_{2j})) \cdot X^{-1}
\]

\[
b_{2j} = \text{RMSE} \left( W_{3j} \cdot (X - P_{j-1}) \right)
\]

(13)

where \( e_j \) feedback the data \( P_j = \phi_3^{-1}(L(e_j)) \cdot (W_{2j})^{-1}, P_0 = 0 \). \( L \) and \( L^{-1} \) represent the normalized function and its reverse function, respectively.

Step 7: For set \( j = j + 1 \), we can determine the jth sub-space features \( \{W_{2j}, b_{2j}\} \) and the \((j + 1)\)th sub-space features \( \{W_{2(j+1)}, b_{2(j+1)}\} \) to be

\[
F_{2j} = \phi_2(X, W_{2j}, b_{2j})
\]

\[
F_{2(j+1)} = \phi_2(X, W_{2(j+1)}, b_{2(j+1)}).
\]

(14)

Step 8: We repeated steps 5–7 \((n - 1)\) times to obtain the sub-space features \( \{F_{21}, \ldots, F_{2n}\} \).

3) Subspace Feature Fusion: Dong et al. [44] demonstrated that if the feature contains correct information, early fusion can be considered as a robust strategy over late fusion by an uncomplicated union of different features into one super vector. We considered two pooling to reduce estimation variance error and bias: average pooling [45] and max pooling, especially max pooling, which is employed in many currently popular models of convolutional neural networks (CNNs), including GoogLeNet [46], VGGNet [47], and AlexNet [48]. Max pooling is also widely used to reduce dimension and feature combination in all types of physiological signals [23, [49], [50]. For the multimodality approach, two different types of input data of EOG \( \{F_{2EOG}^{1}, \ldots, F_{2EOG}^{m}\} \) and EEG \( \{F_{2EEG}^{1}, \ldots, F_{2EEG}^{m}\} \) were obtained from the sub-space feature extraction layer. The feature vectors of EEG and eye movements are directly concatenated into a larger feature vector as the inputs. Then, we found feature fusion to be

\[
\hat{F} = g \left( F_{EOG}, F_{EEG} \right) \left\{ \text{mean}_m \left( F_{EOG}, F_{EEG} \right), \text{max}_m \left( F_{EOG}, F_{EEG} \right) \right\}
\]

(15)

where \( g \) is a combination operator.

According to our previous studies [51], [52], given distinct \( \mathbb{N} \) samples \( \{(x_1, Y) \}_{n=1}^{\mathbb{N}}, X \in \mathbb{R}^n, Y \in \mathbb{R}^m \), if the following conditions are met:

\[
W_2 = \phi_2^{-1}(L(e_{n-1})) \cdot X'(K / I + XX')^{-1}
\]

\[
= -(\log(1/L(e_{n-1}) - 1) \cdot X'(K / I + XX')^{-1}
\]

\[
b_2 = \sum (W_2 \cdot X - \phi_2^{-1}(L(e_{n-1}))) / N
\]

\[
= \sum (W_2 X + \log(1/L(e_{n-1}) - 1)) / N
\]

\[
W_3 = \left( e_{n-1}, L^{-1}(F_2) \right) / L^{-1}(F_2)
\]

the equation \( \lim_{n \to \infty} \left| Y - \left( L^{-1}(\phi_2(W_{21}, X + b_{21})) \cdot W_{31} + \cdots + L^{-1}(\phi_2(W_{2j}, X + b_{2j})) \cdot W_{3j} \right) \right| = 0 \) holds. Both the input and output weights of the proposed method are shown to have the smallest norm among all the least-squares methods.

4) Feature Combination: Since the data have been divided into five batches, we obtained the feature of the 1st batch \( \{F_{21}^{1}, \ldots, F_{2n}^{1}\} \) by performing steps 1–8 one time.

Step 9: We repeated steps 1–8 \((5 - 1)\) times to obtain the entire sub-space feature \( \{F_{21}^{1} + F_{21}^{2} + F_{21}^{3} + F_{21}^{4} + F_{21}^{5}\} \).

C. Regression for Vigilance Estimation

Yang and Wu [28] indicated that mixed neurons play a vital role in the coding and functioning of our brains. By recasting sub-space features into the mapping space, relevant brain signals can be extracted by these features while generating complex and stable behaviors. This process, from dimension reduction and signal reconstruction to feature fusion, as illustrated in Fig. 6, shows the learning structures and dimensions related to the above-mentioned biological evidence. We used such a model to process signals recognition. The entire data of one experiment were separated into five sessions for evaluation, and after fusing all sessions’ features, we used five-fold cross-validation to evaluate the performance. The value of the continuous output data \( y \) in the range of \( 0-0.35, 0.36-0.70, \) and \( 0.71-1 \) indicates the awake state, the fatigue state, and the drowsy state, respectively.

The mean root-mean-square error (RMSE) and the mean correlation coefficient (COR) are used to quantitatively assess the extent of vigilance like quantitative testing of alcohol in the blood. RMSE and COR usually reflect the squared error and linear relationship between the observed and predicted values, respectively. The range of the COR value is \([-1, 1]\), where \(-1, 0, 1\) represent the most disagreement, lack of linear relationship, and the most agreement, respectively. The formulas are

\[
\text{RMSE}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2 / n}
\]

\[
\text{COR}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}
\]

(17)

where \( x = (x_1, x_2, \ldots, x_n)^T \) and \( y = (y_1, y_2, \ldots, y_n)^T \) represent the observed values and the predicted values, respectively, while \( \bar{x} \) and \( \bar{y} \) represent the average value of \( x \) and \( y \), respectively. In short, the lower the RMSE value, and the higher the COR value, the higher the accuracy of the predicted regression.

Analysis of variance (ANOVA) [53] is not only used to study the statistical models and their associated estimation procedures between groups but also within a group. We used ANOVA to assess the statistical significance of the final experimental results. According to Fisher’s F statistic [54], the observed \( F \)-value can be calculated with the original data; the empirical frequency distribution of a new \( F \)-value—that is, \( F^* \)-value—can be obtained through the labels permuted randomly, which belong to a particular group. Thus, the \( F \)-value
Fig. 6. Proposed method with \( n \)-channel autoencoder network and each channel comprising an \( l \)-layer structure.

TABLE VII

<table>
<thead>
<tr>
<th>Methods</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELM</td>
<td>Grid search in ( 2^{10} \sim \cdots \sim 10^{10} ); 1000 hidden neurons are used.</td>
</tr>
<tr>
<td>B-ELM</td>
<td>Grid search in ( 2^{10} \sim \cdots \sim 10^{10} ); 1000 hidden neurons are used.</td>
</tr>
<tr>
<td>SVR</td>
<td>Grid search in ( 2^{10} \sim \cdots \sim 10^{10} ); Linear Kernel.</td>
</tr>
<tr>
<td>CCRF</td>
<td>Regularization hyper-parameters: ( \alpha_k = 10^{0.1} ); Vertex features ( K_1 = (10,20,30); K_2 = 1 ); The sequence length ( n = 7 ); If ( i^{th} ) and ( j^{th} ) neurons are adjacent, ( S^{(k)} = 1 ); otherwise, ( S^{(k)} = 0 ).</td>
</tr>
<tr>
<td>CCN</td>
<td>Regularization hyper-parameters: ( \beta_k = 10^{3} \sim 2 \sim 1.0 ); Vertex features ( K_1 = (10,20,30); K_2 = 1 ); The sequence length ( n = 7 ); If ( i^{th} ) and ( j^{th} ) neurons are adjacent, ( S^{(k)} = 1 ); otherwise, ( S^{(k)} = 0 ).</td>
</tr>
<tr>
<td>DNN(_{SN})</td>
<td>( C_1 = 2^{10} \sim \cdots \sim 10^{10} ); ( C_2 = 2^{10} \sim \cdots \sim 10^{10} ); Three subnetwork neurons are used, each of which contains 500 hidden neurons.</td>
</tr>
<tr>
<td>Outs(_{ap})</td>
<td>For single modality: ( C_1 = 2^{10} \sim \cdots \sim 10^{10} ); ( C_2 = 2^{10} \sim \cdots \sim 10^{10} ); Three subnetwork neurons are used, each of which contains 400 hidden neurons.</td>
</tr>
<tr>
<td>Outs(_{ap})</td>
<td>For multi-modality: ( C_1 = 2^{10} \sim \cdots \sim 10^{10} ); ( C_2 = 2^{10} \sim \cdots \sim 10^{10} ); Three subnetwork neurons are used, each of which contains 400 hidden neurons.</td>
</tr>
</tbody>
</table>

from the \( F \) statistic based upon \( F^* \) is the probability of the true null hypothesis (\( H_0 \)), which is calculated as the proportion of the \( F^* \) that is greater than or equal to \( F \), as follows:

\[
P = \frac{\text{Numbers}(F^* \geq F_{\text{observed}})}{\text{Total Numbers}(F^*)}.
\]

IV. EXPERIMENTAL VERIFICATION

We tested all of the algorithms outlined in this section with MATLAB 2019a with 64-GB memory. The valuation of parameters can be tuned in every step of the experiment, and Table VII shows the details.

A. Single Modality

1) Using Forehead EOG: We compared the regression models that are commonly utilized with EOG features for vigilance estimation: ICA\(_f\), MINUS\(_f\), ICA\(_f\)-MIN\(_fh\), ELM [55], bidirectional-ELM (B-ELM) [56], DNN\(_{SN}\) [23], and the proposed method. The three different features of EOG\(_{fh}\)-ICA, EOG\(_{fh}\)-ICA, EOG\(_{fh}\)-MINUS, and EOG\(_{fh}\)-MINUS were extracted by the MINUS and ICA separation strategies. We then obtained three types of EOG features, including ICA\(_f\), ICA\(_f\)-MIN\(_fh\), and MINUS\(_f\).

The performance of these regression models on three types of EOG\(_f\) features is shown in Fig. 7, including the mean RMSE/COR (\( \text{RMSE}_m / \text{COR}_m \)) and \( \text{RMSE}_\sigma / \text{COR}_\sigma \). Here, \( \sigma \) represents the SD. The ICA and MINUS methods have been shown to have the advantage of regressing high-dimensional features using the big training dataset. The mean RMSE/COR of the ICA\(_f\), ICA\(_f\)-MIN\(_fh\), and MINUS\(_f\) is 0.16/0.48, 0.12/0.78, and 0.13/0.72, respectively. It is found that the blink and saccade components can be easily detected by the ICA\(_f\)-MIN\(_fh\) separation method from the EOG\(_f\) signal, which shows a better performance, and we use its performance as the benchmark. The ELM model is frequently used in regression and has an effective and trustworthy performance. ELM is inherently a single-layer feedforward neural network, meaning it can recognize multiple EOG features. The mean RMSE/COR of the ICA\(_f\), ICA\(_f\)-MIN\(_fh\), and MINUS\(_f\) is 0.16/0.48, 0.12/0.78, and 0.13/0.72, respectively. The blink and saccade components can be easily detected by the ICA\(_f\)-MIN\(_fh\) separation method from the EOG\(_f\) signal, which shows a better performance, and we use its performance as the benchmark. The ELM model is frequently used in regression and has an effective and trustworthy performance. ELM is inherently a single-layer feedforward neural network, meaning it can recognize multiple EOG features. The mean RMSE/COR of the ELM using EOG\(_f\) features is improved to 0.13/0.67, 0.13/0.72, and 0.13/0.73, respectively. DNN\(_{SN}\) is another strong learning method that improves the overall performance of a series of regressors. We observe that the mean RMSE\(_m / \text{COR}_m\) is greatly improved to 0.12/0.72, 0.11/0.79, and 0.11/0.78, respectively. Moreover, the performance of our single-modality algorithm with EOG\(_f\) notably improved to 0.11/0.79 (\( p < 0.01 \)/\( p < 0.01 \), ANOVA), 0.10/0.83 (\( p < 0.01 \)/\( p < 0.01 \), ANOVA), and 0.10/0.80 (\( p < 0.01 \)/\( p < 0.01 \), ANOVA),
Fig. 7. Performance using different EOG features. \( \sigma \) represents its STD.

respective. The mean performances of all compared single-modal methods using EOG are listed in Table VIII, and the proposed method for obtaining the best mean performance of RMSE\(_m\)/COR\(_m\) is 0.10/0.81, which far outperformed other single methods. Our strategy should be seen as a good technique for detecting blink, glances, and fixation components of vigilance.

2) Using EEG: The performance of these regressors using six EEG signals—(EEG\(_f\), EEG\(_t\), EEG\(_p\), EEG\(_g\), EEG\(_f\), and EEG\(_p\))—is shown in Figs. 8 and 9. The mean performances of all compared single-modal methods using EEG are listed in Table IX. Here, \( f, t, \) and \( p \) represent forehead, temporal, and posterior, respectively. For example, EEG\(_{g,f}\) represents EEG signals gathered from the forehead site of the brain, which are separated from the five frequency bands method.

After using two approaches—MA and LDS filtering—four different features of each EEG signal were extracted: DE-MA, DE-LDS, PSD-MA, and PSD-LDS. We found that the DE feature has reliably recognized EEG patterns between low and high-frequency energy due to the comparison regression models, which include ICA, ELM, B-ELM, DNN\(_S\), and respectively. The mean performances of all compared single-modal methods using EOG are listed in Table VIII, and the proposed method for obtaining the best mean performance of RMSE\(_m\)/COR\(_m\) is 0.10/0.81, which far outperformed other single methods. Our strategy should be seen as a good technique for detecting blink, glances, and fixation components of vigilance.

2) Using EEG: The performance of these regressors using six EEG signals—(EEG\(_f\), EEG\(_t\), EEG\(_p\), EEG\(_g\), EEG\(_f\), and EEG\(_p\))—is shown in Figs. 8 and 9. The mean performances of all compared single-modal methods using EEG are listed in Table IX. Here, \( f, t, \) and \( p \) represent forehead, temporal, and posterior, respectively. For example, EEG\(_{g,f}\) represents EEG signals gathered from the forehead site of the brain, which are separated from the five frequency bands method.

After using two approaches—MA and LDS filtering—four different features of each EEG signal were extracted: DE-MA, DE-LDS, PSD-MA, and PSD-LDS. We found that the DE feature has reliably recognized EEG patterns between low and high-frequency energy due to the comparison regression models, which include ICA, ELM, B-ELM, DNN\(_S\), and respectively. The mean performances of all compared single-modal methods using EOG are listed in Table VIII, and the proposed method for obtaining the best mean performance of RMSE\(_m\)/COR\(_m\) is 0.10/0.81, which far outperformed other single methods. Our strategy should be seen as a good technique for detecting blink, glances, and fixation components of vigilance.

2) Using EEG: The performance of these regressors using six EEG signals—(EEG\(_f\), EEG\(_t\), EEG\(_p\), EEG\(_g\), EEG\(_f\), and EEG\(_p\))—is shown in Figs. 8 and 9. The mean performances of all compared single-modal methods using EEG are listed in Table IX. Here, \( f, t, \) and \( p \) represent forehead, temporal, and posterior, respectively. For example, EEG\(_{g,f}\) represents EEG signals gathered from the forehead site of the brain, which are separated from the five frequency bands method.
and the proposed method; using the DE feature therefore appears to have an effect consistent with that found in previous studies [57].

We observed that ICA regressor EEG-based vigilance estimation has a promising performance and the better RMSE_{m}/COR_{m} of EEG_{f}, EEG_{t}, and EEG_{p} which are
that the DNN SN approach has a remarkable performance, and has performed well on each of the EEG features. We noticed performance as the benchmark. Similarly, the ELM regressor achieves the best results and far outperforms other methods. Compared to other fusion strategies, the proposed method uses input features extracted by the calculated decoding layer of the multichannel autoencoder model and uses the subnetwork neurons of multilayer regression models to extract subspace features and fusion subspace features, based on early fusion, which could allow it to obtain a nearly 12% boost.

The performance of the proposed multimodality method with average pooling and max pooling is 0.08/0.88 ($p < 0.01/p < 0.01$, ANOVA) and 0.08/0.89 ($p < 0.01/p < 0.01$, ANOVA), respectively. Furthermore, from training the original signal to displaying detection, the proposed method only takes 3 s to achieve good performance (see Fig. 10), which is nearly ten times faster than other methods. The robustness of the proposed method is proven through the lowest RMSE$_m$, the highest COR$_m$, and the lowest time cost, obtained with two pooling types in subspace feature combination. All experimental results of the compared multimodality algorithms perform better than the single-modality method, and the proposed multimodal method performs the best of all.

### Table X

**Experimental Results for Multimodality. The Best Results Are Bolded**

<table>
<thead>
<tr>
<th>Methods</th>
<th>ELM</th>
<th>B-ELM</th>
<th>SVR</th>
<th>CCRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE$_m$</td>
<td>0.11</td>
<td>0.12</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>COR$_m$</td>
<td>0.78</td>
<td>0.70</td>
<td>0.83</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Comparing to other states, eye movement alone is not enough to develop a standing information processing mechanisms [62]. We then used mixed features with complementary characteristics. The proposed multimodality method shows strong performance in identifying the vigilance states of our brain activity and proves to work better than other state-of-the-art single modality and multimodality approaches.

Although eye movement is a promising indicator of arousal states, eye movement alone is not enough to develop a

![Fig. 10. Computational time analysis.](image-url)

In short, whether using EOG or EEG, the proposed method achieves the best results and far outperforms other methods.

### B. Multimodality

We used the complementarity characteristic between EOG and EEG signals to test various multimodal regression methods with features fusion to assess levels of vigilance: ELM, B-ELM, autoencoder-ELM (AE-ELM) [52], support vector regression (SVR) [59], continuous conditional random field (CCRF) [60], continuous conditional neural field (CCNF) [61], DNN$_{SN}$, and the proposed method with two pooling types: 1) Ours$_{ap}$ and 2) Ours$_{mp}$. Table X shows all experimental results and Ours$_{ap}$ and Ours$_{mp}$ are the proposed method, with average pooling and max pooling, respectively.

The mean RMSE/COR of ELM was significantly improved to 0.11/0.78, for which performance is much better than for its single modality. SVR achieves the COR value of 0.83, which shows that it is a popular and robust regression method in machine learning.

In addition, we can also observe that the RMSE$_m$/COR$_m$ of the CCNF and CCRF with temporal dependency is 0.10/0.84 and 0.09/0.85, respectively, marking a great improvement in performance, which proves its ability to predict continuous vigilance levels. The convolution parameter errors produce the deviation in the mean estimates, which can be reduced by early fusion with the max pooling used in the DNN$_{SN}$ model.

The performance of the DNN$_{SN}$ model obviously improved to 0.09/0.85 and demonstrated its effectiveness.

In this article, we proposed a novel multilayer network structure for vigilance estimation, MAE-MELM$_{int}$, which is composed of the multichannel autoencoders with subnetwork neurons for dimensionality reduction and signal reconstruction. Moreover, compared with other methods of feature selection, the training of our system achieves higher learning accuracy. Simultaneously, the higher efficiency of decoding the brain signals can better identify the specific relationship between the brain activity and cognitive state, while providing evidence and support to aid in decoding brain states and understanding information processing mechanisms [62].

### V. Conclusion

In this article, we proposed a novel multilayer network structure for vigilance estimation, MAE-MELM$_{int}$, which is composed of the multichannel autoencoders with subnetwork neurons for dimensionality reduction and signal reconstruction. Moreover, compared with other methods of feature selection, the training of our system achieves higher learning accuracy. Simultaneously, the higher efficiency of decoding the brain signals can better identify the specific relationship between the brain activity and cognitive state, while providing evidence and support to aid in decoding brain states and understanding information processing mechanisms [62].
robust vigilance estimation model. Eye movement can be intentionally controlled by subjects, which causes degraded performance in prediction. Recently, the robustness of intelligent systems based on machine learning has drawn great attention [63]. Intentional eye movements could be considered as adversarial examples in comparison with spontaneous eye movements. However, we can leverage additional information from EEG to differentiate spontaneous and intentional eye movements. The changes in brain activities contribute to early warnings of reduced vigilance. How to improve the robustness of multimodal vigilance estimation systems need further systematic investigation.

We noticed that the physiological signal varies from person to person. If we obtain more experimental samples and select a larger age range, we can also verify the effectiveness of this model, undoubtedly providing more convincing results. Due to research funding and time constraints, however, all subjects were students recruited from the university campus, with a relatively narrow age range. Using experimental data, we could work to better understand the relationship between age and this model. This is the focus of our future work. Meanwhile, we would like to propose an efficient general method of converting the tabular signals into 2-D shape signals. By doing so, the CNNs, such as ResNet and DenseNet, could be directly combined to improve performance.

REFERENCES


Yimin Yang (Senior Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent systems from the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2013. From 2014 to 2018, he was a Postdoctoral Fellow with the University of Windsor, Windsor, ON, Canada. He is currently an Assistant Professor with the Computer Science Department, Lakehead University, Thunder Bay, ON, Canada. His current research interests include artificial neural networks, signal processing, and robotics.

Dr. Yang was a recipient of the Outstanding Ph.D. Thesis Award of Hunan Province and the Outstanding Ph.D. Thesis Award Nominations of Chinese Association of Automation, China, in 2014 and 2015, respectively. He is a Program Committee Member of some international conferences. He is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and Neurocomputing. He has been serving as a reviewer for many international journals of his research field and a Guest Editor of multiple journals.

Hui Zhang (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in pattern recognition and intelligent systems from Hunan University, Changsha, China, in 2004, 2007, and 2012, respectively. He is currently a Professor with the College of Robot, Hunan University, where he is also the Deputy Director of the National Engineering Laboratory of Robot Visual Perception and Control Technology. His research interests include machine vision, deep learning, and defect detection.

Wei-Long Zheng (Member, IEEE) received the bachelor’s degree in information engineering from the Department of Electronic and Information Engineering, South China University of Technology, Guangzhou, China, in 2012, and the Ph.D. degree in computer science from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2018. He is a Research Fellow with the Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. His research focuses on affective computing, brain–computer interaction, machine learning, and pattern recognition.

Dr. Zheng received the IEEE TRANSACTIONS ON AUTONOMOUS MENTAL DEVELOPMENT Outstanding Paper Award from the IEEE Computational Intelligence Society in 2018.

Bao-Liang Lu (Senior Member, IEEE) received the B.S. degree in instrument and control engineering from the Qingdao University of Science and Technology, Qingdao, China, in 1982, the M.S. degree in computer science and technology from Northwestern Polytechnical University, Xi’an, China, in 1989, and the Dr.Eng. degree in electrical engineering from Kyoto University, Kyoto, Japan, in 1994. He was with the Qingdao University of Science and Technology from 1982 to 1986. From 1994 to 1999, he was a Frontier Researcher with the Bio-Mimetic Control Research Center, Institute of Physical and Chemical Research (RIKEN), Nagoya, Japan, and a Research Scientist with the RIKEN Brain Science Institute, Wako, Japan, from 1999 to 2002. Since 2002, he has been a Full Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests include brain-like computing, neural networks, machine learning, brain–computer interaction, and affective computing.

Prof. Lu received the IEEE TRANSACTIONS ON AUTONOMOUS MENTAL DEVELOPMENT Outstanding Paper Award from the IEEE Computational Intelligence Society in 2018. He is currently a Board Member of the Asia Pacific Neural Network Society (APNNS, previously APNNA) and the Steering Committee member of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING. He was the President of the Asia Pacific Neural Network Assembly (APNA) and the General Chair of the 18th International Conference on Neural Information Processing in 2011. He is currently an Associate Editor of the IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENT SYSTEMS.