Temporal-Noise-Aware Neural Networks for Suicidal Ideation Prediction Using Physiological Data

Niqi Liu, Fang Liu, Xinxin Du, Yezhi Shu, Xu Liu, Lan Wang, Guozhen Zhao, Wenting Mu, and Yong-Jin Liu, *Senior Member, IEEE*

***Abstract*—The robust generalization of deep learning models in the presence of inherent noise remains a significant challenge, especially when labels are ambiguous due to their subjective nature and noise is indiscernible in natural settings. In this article, we address a specific and important scenario of monitoring suicidal ideation (SI), where time-series data, such as galvanic skin response (GSR) and photoplethysmography (PPG), are susceptible to such noise. Current methods predominantly focus on image and text data or address artificially introduced noise, neglecting the complexities of natural noise in time-series analysis. To tackle this, we introduce a novel neural network model tailored for analyzing noisy physiological time-series data, named DBN_ConvNet, which integrates advanced encoding techniques with confidence learning training to enhance prediction performance. Another main contribution of our work is the collection of a specialized dataset of GSR and PPG signals derived from real-world environments for SI prediction. By employing this dataset, our DBN_ConvNet achieves a prediction accuracy of 76.67% and an F1 score of 0.74 in a binary classification task, outperforming state-of-the-art methods. Furthermore, comprehensive evaluations have been conducted on three other well-known public datasets with artificially introduced noise to test the DBN_ConvNet's capabilities rigorously. These tests consistently demonstrated DBN_ConvNet's superior performance by achieving an improvement of more than 10% in both accuracy and F1 score compared to the baseline methods.**

***Index Terms*—Learning with noise, peripheral physiological signals, suicidal ideation prediction.**

Received 22 October 2023; revised 23 June 2024 and 26 December 2024; accepted 26 December 2024. This work was supported in part by the Natural Science Foundation of China under Grant U2336214, in part by the National Key Research and Development Plan under Grant 2018YFC0831001, and in part by the China Postdoctoral Science Foundation under Grant 2024M751591. *(Niqi Liu and Fang Liu contributed equally to this work.) (Corresponding authors: Guozhen Zhao; Yong-Jin Liu.)*

Niqi Liu is with the CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing 100101, P. R. China, also with BNRist, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China, and also with the MOE-Key Laboratory of Pervasive Computing, Beijing 100084, China (e-mail: lnq22@mails.tsinghua.edu.cn).

Fang Liu is with the School of Information and Communication Engineering, Communication University of China, Beijing 100024, China (e-mail: fangliu@cuc.edu.cn).

Xinxin Du, Yezhi Shu, and Yong-Jin Liu are with BNRist, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China, and also with the MOE-Key Laboratory of Pervasive Computing, Beijing 100084, China (e-mail: duxx@tsinghua.edu.cn; shuyz19@mails; liuyongjin@tsinghua.edu.cn).

Guozhen Zhao is with the CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing 100101, P. R. China, and with the University of Chinese Academy of Sciences, Beijing 100049, P. R. China, and also with the Multimodal Sensing and Computing Laboratory, Beijing 100083, P. R. China (e-mail: zhaogz@psych.ac.cn).

Xu Liu and Lan Wang are with the Multimodal Sensing and Computing Laboratory, Beijing 100083, P. R. China (e-mail: liuxuhenu@gmail.com; lan.wang.98@outlook.com).

Wenting Mu is with the Department of Psychological and Cognitive Sciences, Tsinghua University, Beijing 100084, China (e-mail: wmu@tsinghua.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TCSS.2024.3523928, provided by the authors.

Digital Object Identifier 10.1109/TCSS.2024.3523928

I. Introduction

SUICIDAL ideation (SI) is a critical public health issue, affecting diverse demographic groups with notable prevalence [1]. Effective detection and timely intervention are crucial for saving lives, yet accurately identifying at-risk individuals poses substantial challenges, especially within high-stress environments and communities marked by mental health stigma [2]. Such environments include colleges [3], aviation sectors [4], and correctional facilities [5], where individuals might conceal their mental states. This concealment significantly complicates the precision of diagnosis, presenting unique challenges in these settings.

Traditional methods for detecting SI typically rely on self-reported data [6] and various behavioral assessments [7], including questionnaires [8], interviews [9], facial expressions [7], vocal patterns [10], and social media activities [11], [12]. However, these methods are significantly compromised by the aforementioned concealment, as individuals can deliberately manipulate their responses and hide their true emotions.

Advances in physiological monitoring provide new avenues for assessing mental health states. Substantial evidence links physiological responses associated with depressive states to SI, offering a reliable pathway for intervention. Existing studies show that depression, anxiety, loneliness, and agitation are key risk factors for SI [13], [14], with the severity of depressive symptoms identified as a significant predictor [15], [16]. Unlike impulsive suicides that can result from sudden stressors, SI related to chronic depression often presents discernible patterns that enhance its predictability. Our study focuses on detecting SI among prisoners arising from such long-term depressive states.

Beyond behavioral expressions, physiological parameters are less susceptible to voluntary manipulation [17] and have been recognized as reliable indicators of specific affective states [18]. Peripheral physiological signals, which reflect activities of the autonomic nervous system (ANS), play a crucial role in emotional responses [19]. Notably, depression is strongly associated with measurable changes in galvanic skin response (GSR) and photoplethysmography (PPG) [20], [21], [22], offering objective markers for early detection that are difficult to camouflage. Additionally, these measures are noninvasive [23], cost-effective, and easily obtained through portable physiological sensors, making them practical for large-scale applications.

Previous research links major depression to reduced emotional responsiveness and altered ANS reactivity [19], [24]. This diminished responsiveness is often observed during emotional tasks, such as viewing film clips [25], [26], with depressed individuals showing less pronounced physiological reactions. Drawing on these findings, our study employs GSR and PPG signal analysis to detect subtle variations in ANS responses triggered by a film clip.

A. Task Definition

Using features extracted from induced GSR and PPG signals, this article approaches the prediction of SI as a binary classification problem, distinguishing between "with SI" and "without SI" labels.

B. Population Context

In this study, we focus on prisoners, a population where the tendency of SI is significantly stronger than in the general population. A Belgian prison survey [27], involving 1203 male prisoners across 15 Flemish prisons, revealed that 43.1% of these respondents reported a lifetime history of SI. Comparable rates of SI were also found in prisons in Australia [28] and China [29]. This situation is exacerbated by the unique psychological and environmental stressors of prison life [30], [31], [32], making this population particularly pertinent to our study.

C. Challenges

1) The utility of physiological signals in real-world applications is often hindered by inherent noise and variability in the data, leading to credibility issues with labels. Positive labels (denoted as "with SI") are assigned to individuals showing overt suicidal behaviors or who have previous suicide attempts, while all others are labeled as negative labels (denoted as "without SI"). Although most negative labels accurately reflect the absence of suicidal ideation, a minority of individuals may conceal their intentions or exhibit subtle signs, resulting in false negatives. 2) Another challenge arises from the technical limitations in existing methodologies. Despite recent advancements in handling noisy time-series data [33], [34], many strategies are borrowed from domains primarily dealing with image data [35] or artificially noised series. These methods often fail to address the unique complexities of noisy physiological time series, which are typically weakly annotated and characterized by natural noise, thus complicating the analysis and interpretation.

D. Main Ideas

This study develops DBN_ConvNet, a novel computational model designed to handle noisy physiological time-series data of GSR and PPG signals. This model leverages semisupervised learning and dynamic noise adaptation through confidence-learning [36], enhancing the reliability of detecting suicidal ideation (SI). Additionally, we have explored existing alternative noisy learning strategies including *MentorNet* [37] and *Coteaching* [38], with EEGNet [39] as the feature extraction backbone, to handle label noise. These strategies focus on the loss of individual samples and are implemented in parallel to assess their efficacy. In comparison, we benchmark our approaches against state-of-the-art networks tackling noise, such as CTW [33], SREA [34], and DivideMix [35] to contextualize their performance in noisy environments. To evaluate the effectiveness of DBN_ConvNet and the noisy learning strategies, we calculate accuracy, F1 score, and Matthews correlation coefficient (MCC) on a clean test set. Furthermore, we assess the model's generalizability by analyzing its performance on publicly available time-series datasets that have been artificially noised.

To summarize, this article makes the following contributions.

1) We introduce DBN_ConvNet, a novel semisupervised learning model that combines a self-supervised encoding module with a supervised convolutional framework specifically designed for noisy data environments. It employs a two-stage training strategy that initially identifies and mitigates label noise via confidence learning, followed by a refinement phase where DBN_ConvNet is retrained on a cleaned dataset, enhancing the reliability and accuracy of SI prediction.
2) We construct a specialized dataset for SI prediction that includes extracted features from physiological signals and corresponding labels. Using an affective reactivity paradigm, we collected GSR and PPG signals from a cohort of 2190 prisoners. With the constructed dataset, we address a significant gap in current research methodologies by successfully pioneering the prediction of SI using peripheral physiological signals in a correctional environment.

Experimental results confirm that our methods outperform state-of-the-art models in SI prediction. Following a subject-independent protocol, DBN_ConvNet achieves an average accuracy of 76.67% (14.17% higher than baselines). In addition, the application of sample-reweighting strategies also proves effective in denoising, improving accuracy by 7.50% and 6.67% for MentorNet and Coteaching, respectively compared to baselines.

II. Related Work

A. Peripheral Physiological Indicators in Depression and Suicidal Ideation Prediction

Previous research has shown the potential of peripheral physiological signals as markers for SI when responding atypically to affective stimuli [26], [40]. These signals are automatic responses from the autonomic nervous system (ANS) to

emotional triggers [41], [42]. Specifically, emotional arousal associated with sympathetic nervous system (SNS) activity can be noninvasively monitored through peripheral physiological measures. GSR and PPG are particularly valuable: GSR captures changes in skin conductance [43], while PPG provides cardiovascular indices such as heart rate variability [44].

Numerous studies have established a correlation between blunted SNS reactivity to psychological stressors and depressive states. Individuals with major depressive disorder (MDD) have demonstrated attenuated heart rate, blood pressure, and GSR during mental stress, irrespective of their subjective perception of stress [45], [46]. Salomon et al. [47] found diminished cardiovascular reactivity to stress exclusively in active MDD cases, suggesting that blunted SNS functioning may be related to current mood rather than a permanent trait. Similar blunted responses to affective stimuli have been documented: emotional films used to induce stress showed reduced physiological reactivity in MDD individuals compared to healthy controls [48]. These patterns highlight emotional disengagement in MDD, potentially explaining associated interpersonal difficulties and empathy deficits [49].

Additionally, even after controlling depressive symptom severity, an inverse correlation between SI and changes in high-frequency heart rate variability (HF-HRV) during mental stress remained significant [40]. This correlation was further supported by studies using sad film clips, where blunted HF-HRV responses differentiated MDD individuals with SI from those without [26].

B. Computational Methods Using Physiological Signals

Traditional machine learning algorithms such as support vector machine (SVM), random forest (RF), naive Bayes (NB), and nearest neighbor (KNN) have commonly been applied to analyze GSR or PPG signals [50], [51], [52]. The field has also seen the development of task-specific deep neural networks [53], [54], [55], [56]. In contrast, EEG-based architectures have demonstrated versatile adaptability and robust generalization across tasks by leveraging central-peripheral nervous system interrelations and compatible data formats. For instance, *EEGNet* [39] features a combination of standard 2D, depthwise, and separable convolutions to simulate bilinear discriminant component analysis for efficient feature integration. *DeepConvNet* [57], designed for general use, incorporates five convolutional layers to provide a versatile framework for feature analysis. *ShallowConvNet* [57] simulates a feature extraction paradigm (filter bank common space mode [58]) for analyzing EEG signals and excels in capturing oscillatory signals. *FBCNet* [59] employs narrowband filters and non-overlapping convolutional windows to capture spatial and temporal features across frequency bands. *DGCNN* [60] and its sparse version [61] dynamically model electrode relationships in a graph, adapting the adjacency matrix for complex multichannel data.

When adapting these for our study, the temporal convolution layers preserve time-series analysis roles, while the spatial convolution layers handle manually extracted features from GSR and PPG signals.

C. Learning With Noisy Labels

During the labeling process for our suicide ideation prediction dataset (Section III-B), some samples with mid-range scores indicated ambiguous SI, potentially introducing noisy labels. To improve data quality and classification reliability, we employ computational methods to automate noise filtering during training.

To learn from noisy labels, researchers have developed robust loss functions [62], [63], utilized regularization techniques [64], and differentiated noisy from clean data based on sample loss [65], [66], [67]. Our approach focuses on identifying and using cleaner samples through effective noise discernment strategies. Correctly labeled samples generally incur smaller losses early in training due to class-dependent characteristics, whereas noisy samples may require multiple training iterations to yield accurate learning. In this context, *confidence learning* quantifies the likelihood of label noise by estimating the joint distribution between provided labels and model predictions, serving as a filtering mechanism before formal training [36]. *MentorNet* introduces a self-adjusting curriculum through a "mentor net" that modulates sample loss weights, allowing the "student net" to use these weights for more accurate learning outcomes [37]. *Coteaching* enhances noise robustness by deploying two classifiers that exchange the least noisy samples for mutual training, thus minimizing error propagation and enhancing the overall noise handling [38].

Besides sample loss methods, self-supervised learning through deep belief networks (DBNs) can also address label deficiencies. DBNs, constructed from layered restricted Boltzmann machines (RBMs) [68], train sequentially in an unsupervised manner. They facilitate bidirectional data flow between pairs of visible and hidden layers until a specific loss threshold is met, enabling the reconstruction of contaminated input data. This architecture minimizes overfitting and enhances noise resilience, functioning independently of labels.

Moreover, recent efforts have been made to manage label noise in the field of time-series data analysis. CTW [33] employs specialized data augmentation to mitigate artificial noise effects. SREA [34] demonstrates strong capabilities in industrial contexts, suggesting potential adaptations for physiological data. Additionally, DivideMix [35] treats noisy label learning as a semisupervised problem, segregating data into presumed clean and noisy subsets to improve training effectiveness.

Despite the advancements above, these methods often rely on artificial noise or are unsuitable for physiological serials, inadequately addressing the complexities of inherently noisy time-series data. Our study seeks to bridge this gap by incorporating advanced noise-handling techniques into our DBN_ConvNet framework, thus refining the analysis of physiological signals with noisy labels.

III. Data Collection

To tackle the issue of predicting SI using peripheral physiological signals with noisy labels, a significant contribution of this study is the creation of a dataset comprising features extracted from GSR and PPG signals obtained via portable

devices during an affective reactivity paradigm, along with corresponding labels.

A. Participants

Participants in this study were male prisoners from a prison in Hunan Province, China, who volunteered following research advertisements within the prison. We excluded individuals who were hospitalized or housed in high-security sectors. The final sample comprised 2190 right-handed prisoners (mean age $=$ 40.96 ± 12.59 years). Among them, nine were serving life sentences, while the remaining participants had sentence lengths ranging from 6 months to 24 years (mean length $= 7.47 \pm$ 4.35 years). The majority of participants were convicted of theft (9.05%), robbery (9.09%), fraud (12.95%), bribery (15.29%), and drug trafficking (15.42%).

Ethical clearance was obtained from the Institutional Review Board of the Institute of Psychology, Chinese Academy of Sciences. Participants voluntarily joined the study, fully informed about its nature, procedures, and their rights, including confidentiality and the option to withdraw at any time.

B. Labels

Participants were labeled as "with SI" (positive, coded as 1) or "without SI" (negative, coded as 0) based on three types of assessments: prisoners' histories of attempted suicide or self-harm, observations made by prison guards, and face-to-face interviews conducted with licensed psychologists. SI ratings were collected from the prison guards responsible for each participant using a one-question survey, which asked: "To what extent do you believe this prisoner is likely to engage in suicidal behavior?" Guards provided their responses on a 10-point Likert scale, ranging from 1 (not at all) to 10 (very likely), based on the prisoner's daily behaviors.

1) Positive Samples: The participants were required to satisfy the following three conditions simultaneously to be assigned a positive label.

(P1) History of attempted suicide or self-harm, or proactive expression of suicidal thoughts.
(P2) Guard's rating score ≥ 6.
(P3) Clinical judgment made by professionals. Participants meeting the first two conditions underwent a face-to-face semistructured interview with pretrained correctional police officers from the Department of Educational Reform and Psychological Correction. The interview further assessed factors contributing to an individual's SI, including will-to-life, degrees of emotional immersion, and despair levels.

Using the above procedure, 30 participants were identified as positive samples and labeled as 1, while all others were labeled as 0 (negative).

2) Negative Samples: While the above conditions ensure the reliability of positive samples, some participants scored in the mid-range or exhibited behavioral camouflage, contributing noise to the 0-labeled group. To address this, 0-labeled samples were split into two subgroups: purified true negatives and those with lower credibility. The purified subgroup had to meet these three conditions simultaneously.

(N1) Participants with no history of attempted suicide or self-harm, and who have never proactively expressed suicidal thoughts.
(N2) Participants who received the lowest rating score (1 on a 10-point Likert scale) concurrently from at least three guards.
(N3) Verification conducted by psychologists. Similar to the true positive samples, participants who met the first two conditions were also subjected to face-to-face semistructured interviews. During these interviews, trained correctional officers provided additional validation of their nonsuicidal states.

This procedure selected a total of 21 true negative samples. Given the relatively low prevalence of suicide in prisons, the larger subgroup, though labeled as "uncertain negative," likely consists predominantly of non-SI cases. This separation helps balance the numbers between true positives and true negatives. Ultimately, the labels are distributed into three groups.

1) *True Positive (TP):* participants with SI, totaling 30 samples.
2) *True Negative (TN):* participants without SI, totaling 21 samples.
3) *Uncertain Negative (UN):* participants with ambiguous assessments or mid-range scores, predominantly without SI, totaling 2139 samples.

C. Stimuli and Apparatus

A 5-min. video clip related to family affection was utilized as the evocative stimulus. The video depicted the reactions of a group of young adults upon witnessing their parents, who were artificially aged by 20 years through special effect makeup. It was selected based on preliminary interviews with correctional officers and a thorough review of the literature, which emphasized the influence of family support on depressive mood and suicidal ideation among incarcerated individuals [69], [70]. A prior validation study with 2805 incarcerated males further confirmed that this video effectively induced sadness and high emotional arousal in most participants. The video clip is available upon request.

Prisoners watched this video during the study, while we concurrently recorded their GSR and PPG signals using a custom wristband (Ergosensing, China). All prisoners wore the wristband on their non-dominant hand. GSR signals were recorded at a sampling rate of 4 Hz and a resolution of 0.01 μS, while PPG signals were obtained at a sampling rate of 100 Hz, utilizing the reflected green light with a wavelength of 532 nm.

D. Procedure

Participants were briefed on the study's purpose, procedures, and their rights—including voluntary participation and confidentiality—before giving informed consent to proceed.

Grouped by cellblocks, after taking seats and wearing wristbands, they underwent two distinct phases: an initial 3-min. quiet sitting period for baseline data collection ("static" phase),

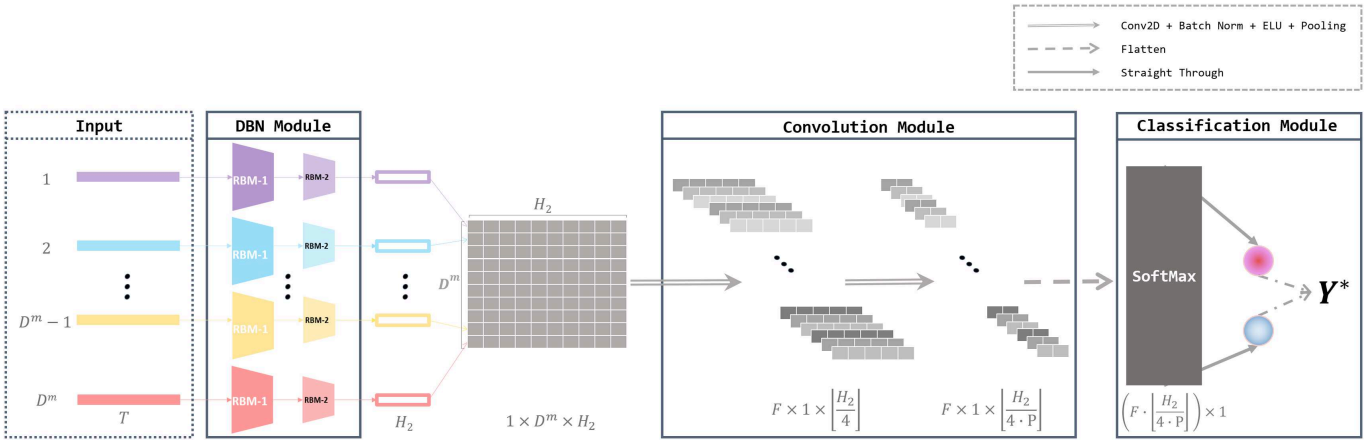

Fig. 1. Proposed DBN_ConvNet model, composed of one DBN module, one convolution module, and one classification module. The input layer of the model is composed of D^m vectors, where D^m denotes the number of extracted features and T represents the dimensionality of each feature vector. H_2 is the number of units in the hidden layer of RBM-2, F is the number of filters (set to 16), $P = \min(H_2//4, 8)$ is a pooling size in the second block of the convolution module, and Y^* is the predicted label.

followed by the "stimulation" phase where they watched the selected video clip.

This setup was crucial for capturing and contrasting baseline electrophysiological responses with those elicited during emotional stimulation, vital for the development of our predictive algorithms. To address this, the amplitude of each timestamp in the "stimulation" phase was adjusted by subtracting the mean value of the "static" phase.

E. Preprocessing

In the preprocessing stage, personal identifiers were replaced by unique codes for access.

1) Feature Extraction and Normalization: *GSR* signals were filtered using a 2nd-order low-pass Butterworth filter with a cutoff frequency of 0.8 Hz. The signals were then cleaned and decomposed into tonic (SCL) and phasic (SCR) components using a Python toolbox [71] and a convex optimization method [72], respectively. Following the decomposition, GSR signals and their components were segmented into 10-s windows with 60% overlap (4-s intervals). Within these windows, 35 time-domain and frequency-domain features were extracted for analysis. *PPG* signals were filtered using a 3rd-order band-pass Butterworth filter with a lower cutoff frequency of 0.6 Hz and a higher cutoff frequency of 5 Hz. These filtered signals were segmented into 20-s windows, each overlapping 80% with the next. Using the HeartPy [73] Python package, features like peak-to-peak interval, root-mean-square of successive differences, pulse width, pulse area, and Shannon entropy were extracted, resulting in 38 distinct features per window. The full details of GSR and PPG features are listed in Appendix A (see the supplementary material).

Signal durations, approximately 300 s, were divided into segments by an interval of 4 s, each yielding 75 samples for GSR and PPG features. Due to minor recording discrepancies, the number of windows varied around 75 ± 5. For consistency, sequences were standardized to 70 windows, resulting in 35×70 and 38×70 feature matrices for GSR and PPG, respectively. These matrices were concatenated to form a 73×70 "FUSE" matrix. Features within each window were normalized using min–max normalization to ensure uniform scale across dimensions.

2) Notations: After normalization, feature sequences and corresponding labels were structured for network processing. Feature sequences for each individual's GSR and PPG, and their fusion are denoted as $\mathcal{V}_{\mathrm{GSR}}[D_{(35)}^{\mathrm{GSR}}, 70]$, $\mathcal{V}_{\mathrm{PPG}}[D_{(38)}^{\mathrm{PPG}}, 70]$, and $\mathcal{V}_{\mathrm{FUSE}}[D_{(73)}^{\mathrm{FUSE}}, 70]$, respectively. Labels $\tilde{y}_i \in \{0, 1\}$ represent the initial label for sample i, potentially incorrect for some UN samples. $\vec{y_i^*} \in \mathbb{R}^2$ denotes the output probabilities, where $\vec{y_i^*}[0]$ and $\vec{y_i^*}[1]$ indicate probabilities of negative and positive labels, respectively. The predicted label $y_i^* \in \{0, 1\}$ matches the coding described in Section III-B

$$y_i^* = \begin{cases} 0, & \vec{y_i^*}[0] \geq \vec{y_i^*}[1] \\ 1, & \text{others} \end{cases} \tag{1}$$

IV. DBN_ConvNet Model

Addressing the technical challenge of robust SI prediction with noisy label data, we introduce DBN_ConvNet, an innovative neural network architecture. This section details DBN_ConvNet, which merges DBN with convolutional modules for noise-aware classification. We also incorporate confidence learning [36] to train DBN_ConvNet effectively in two stages. The subsequent subsections elaborate on the network's structure and training methodology.

A. Architecture

As shown in Fig. 1, DBN_ConvNet consists of three parts: a DBN module, a convolution module, and a classification module.

For *the DBN module*, each feature channel is independently processed by a corresponding DBN instance. As a result, for modality $m \in \{\mathrm{GSR}, \mathrm{PPG}, \mathrm{FUSE}\}$, there are $D^m = 35$, 38,

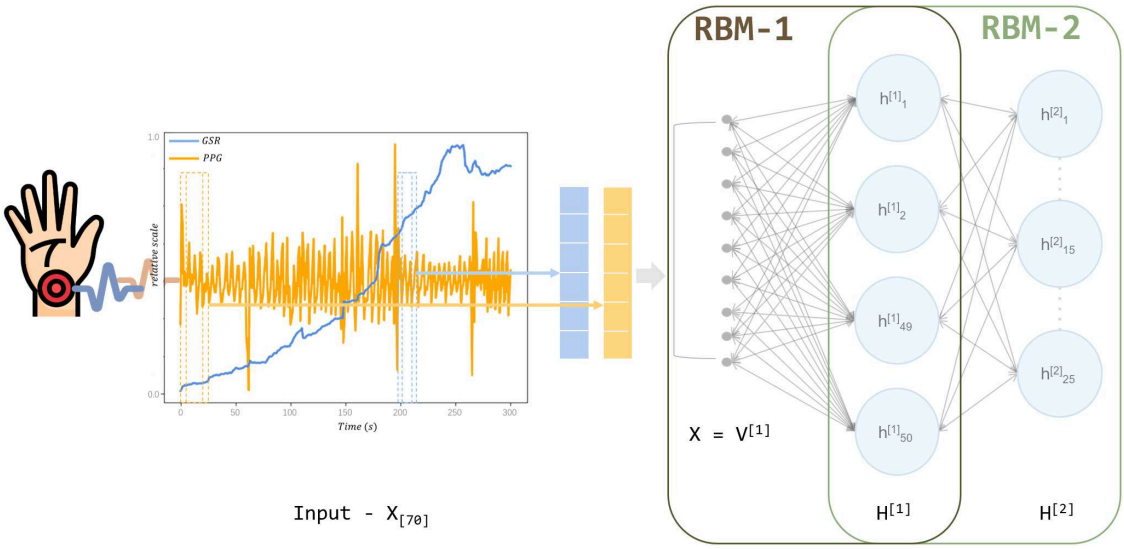

Fig. 2. Architecture of the DBN module of the proposed DBN_ConvNet. RBM-1 and RBM-2 constitute the module, consisting of one 70-unit input layer ($V^{[1]}$), one 50-unit hidden layer ($H^{[1]}$), and one 25-unit output layer ($H^{[2]}$). For each feature extracted from the preprocessed PPG (or GSR) data, the windows are formatted into a 70-dim vector as the DBN input X.

and 73 DBN instances, respectively. Specifically, the input matrix is divided into D^m subsequences based on the feature dimension, with each sequence forming a T-dimensional vector $\vec{v_f}$ for the fth feature ($f = 1, 2, ..., D^m$). These vectors are then transformed into encoded vectors $\vec{h_f}$, as defined in (2), using the weight matrices ($\mathcal{W}_{f1}$, $\mathcal{W}_{f2}$) and biases (b_{f1}, b_{f2}) of the RBMs

$$\vec{h_f} = \mathcal{W}_{f2}(\mathcal{W}_{f1}\vec{v_f} + b_{f1}) + b_{f2}. \tag{2}$$

Fig. 2 presents the details of each DBN instance: two cascading RBMs with $T = 70$ input units (consistent with the number of windows), 50 units in the first hidden layer, and 25 units in the last embedding layer.

This encoding process not only compresses the input data but also minimizes reconstruction loss [see (5)], ensuring the retention of critical time-series data components and effectively handling the variations among the features.

The convolution module concatenates $\vec{h_f}$ back into a matrix $\mathcal{H}$ of size $[D^m, 25]$, aggregating the compact representations from different feature channels. Subsequently, a set of 2-D filters (with dimensions $D^m \times 1$) are applied to these features to learn feature weights and reduce the feature dimension.

The classification module, consisting of a fully connected layer and a softmax layer, outputs two probabilities, one for each label type. The final label for a sample is assigned based on the higher probability between these two outputs.

The detailed data dimensionality for each module is listed in Appendix C (Table A2) (available online).

B. Semisupervised Training With Confidence Learning

We employ a two-stage training approach integrated with confidence learning (CL) (see the purple arrows in Fig. 3). *The first stage* involves generating rough predictions for label confidence assessment and noise identification. *In the second stage*, the training dataset is refined by filtering out the noise identified through CL, thereby enhancing DBN_ConvNet's prediction accuracy. For both of these stages, the DBN_ConvNet is independently initialized and trained with the same hyperparameters and validation protocols, with the only difference lying in the database used.

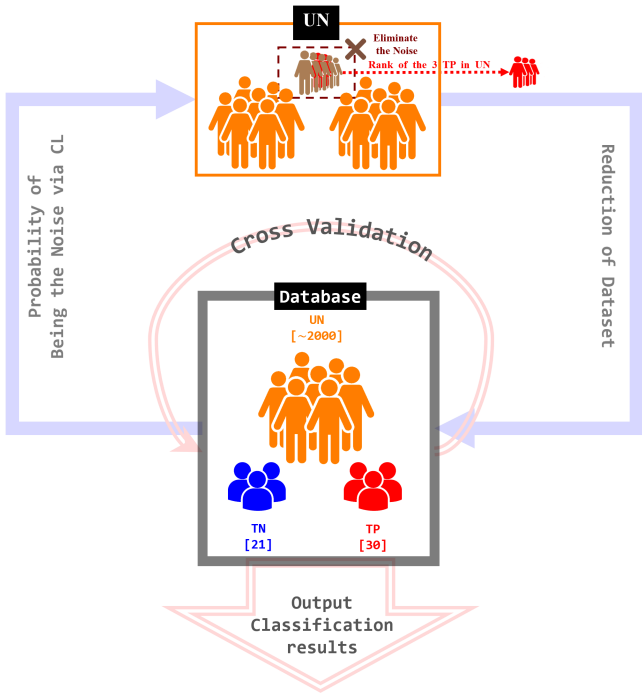

Fig. 3. Validation pipeline of DBN_ConvNet. The pink semicircuit represents a cross-validation stage, which contains ten folds (see Fig. 4 for one fold). The purple arrows draw the whole circuit of our CL-aided procedure.

The pseudocode of the entire two-stage training process is summarized in Algorithm 1, where steps 5 constitute the

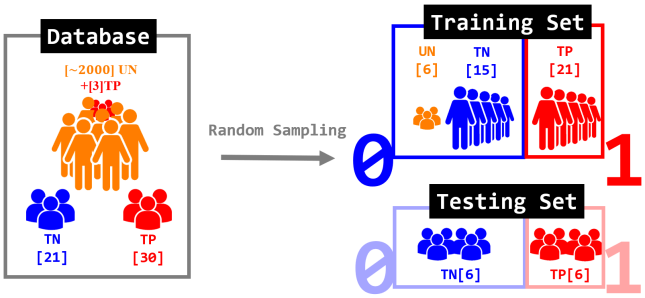

Fig. 4. Visualization of the partition of each fold in cross-validation.

Algorithm 1: Training and Predicting Process of DBN_ConvNet.

Require:

The set of true positive samples, *TP*;
The set of true negative samples, *TN*;
The set of original uncertain negative samples, UN_1;
DBN_ConvNet with parameter θ;

Ensure:

Likelihood of being with suicidal ideation for each sample, $\vec{y^*}_i$.

1: **while** not all $\vec{y^*}_i^1$ are calculated **do**
2: Initialize θ randomly;
3: Select samples from *TP*,*TN*, and UN_1 to build current dataset as Table I;
4: Unsupervised training of the DBN module;
5: Supervised training of the entire network until converged;
6: Credit $\vec{y^*}_i^1$ of samples in the current test set to the total accumulated average.
7: **end while**
8: Calculate the joint distribution matrix $Q_{y^*,\tilde{y}}$ and filer noisy samples from UN_1 according to PBNR method to get UN_2 with higher purity.
9: Repeat step 1-7 and replace UN_1 and $\vec{y^*}_i^1$ by UN_2 and $\vec{y^*}_i^2$.
10: **return** $\vec{y^*}_i^1$ and $\vec{y^*}_i^2$;

training process and step 6 makes the prediction for each stage. Step 8 involves confidence learning, where $\vec{H}_i^1$ represents the inferred result in the first stage, and $\vec{H}_i^2$ represents the inferred results in the second stage for sample i.

1) Training Process: DBN_ConvNet's training integrates a self-supervised phase followed by a supervised phase in both of its stages.

During the self-supervised phase, data undergoes bidirectional processing in each RBM unit of the DBN module across three epochs, and the parameters are updated using the reconstruction loss $\mathcal{L}_{\text{con}}$ [see (5)], defined as follows:

$$\vec{h_{fj}}^k = \mathcal{W}_{fj}\vec{v_{fj}}^k + b_{fj} \tag{3}$$

$$\vec{v_{fj}}^{k+1} = \mathcal{W}_{fj}{}^T\vec{h_{fj}}^k + b^*_{fj} \tag{4}$$

TABLE I
SAMPLE COMPOSITION OF THE DATASET IN EACH FOLD

	# of TP	# of TN	# of UN
Training dataset	21	15	6
Test dataset	6	6	0

Note: "#" Denotes the number of samples.

$$\mathcal{L}_{\text{con}}^{fjk} = \|\vec{v_{fj}}^k - \vec{v_{fj}}^{k+1}\|_1 \tag{5}$$

where $k \in \{0, 1, 2\}$ is the iteration index, $f = 1, 2, ..., D^m$ is the DBN index for the fth feature, $j \in \{1, 2\}$ is the RBM index, $\mathcal{W}_{fj}$ is the weight matrix of *RBM-j* of the fth DBN ($\mathcal{W}_{fj}{}^T$ is its transpose), and b_{fj}, b^*_{fj} are the forward and backward biases of *RBM-j* of the fth DBN.

During the supervised phase, the network switches to a unidirectional training mode. Data sequentially passes through the DBN, convolution, and classification modules. The cross-entropy loss [defined in (6)] is computed right after the softmax layer

$$\mathcal{L} = -\vec{y_i^*}[\tilde{y}_i] + \log\sum_{c=0}^{1} e^{\vec{y_i^*}[c]}. \tag{6}$$

2) Noise Filtering by Confidence Learning: DBN_ConvNet incorporates confidence learning (CL) [36] as a crucial step in its training process, encompassing probability prediction, joint distribution computation, and noise filtering.

Each training stage begins with a selection of samples, including TP, TN, and approximately 2000 UN samples (as defined in Section III-B). After the ten-fold cross-validation, model-inferred probabilities are used to gauge *label confidence* ($\vec{y_i^*}[\tilde{y}_i]$) for each sample, aiding in identifying potential noise.

The process employs the PBNR method [36] to quantify the amount of noise ($\mathcal{N}_{\text{noise}}$) in the UN sample set. Samples with the lowest *label confidence* scores are then excluded, refining the training set for the subsequent stage. This CL-based approach is anticipated to enhance DBN_ConvNet's performance, especially in the second training stage with a more purified dataset. The experimental results in Section VI-C validate this point.

Readers are referred to Appendix B (available online) for full implementation details about CL.

V. SUICIDAL IDEATION PREDICTION SYSTEM

Our system assesses the probability of suicidal ideation (SI) in prisoners, using feature matrices detailed in Section III-E. To address label uncertainty, we employ two strategies: sample-reweighting with MentorNet [37] and Coteaching [38] using EEGNet [39], and direct noise management through DBN_ConvNet (Section IV) with confidence learning (CL) [36]. These approaches lead to three classification methods, whose training and operational details will be further elaborated in the following sections.

A. Implementation of the Classification Methods

The implementation details of the three models in the prediction system are as follows.

1) *MentorNet-EEGNet*: This model employs EEGNet as the *StudentNet* and retains the original structure of *MentorNet* described in [37]. During the training phase, the sample loss and iteration-related elements are structured and fed into MentorNet.
2) *Coteaching-EEGNet*: This model utilizes two instances of EEGNet as the subnets, following the approach in [38]. During the testing phase, the instance that performs better on the training set is selected as the classifier.
3) *DBN_ConvNet*: This is the deep neural network model proposed in Section IV. It performs the classification task in two stages, utilizing the CL method.

The classification modules for all methods employ a 2-unit output layer as described in Section IV-A, which outputs the probability vector $\vec{y_i^*}$. Each classifier is trained with the Adam optimizer using cross-entropy loss $\mathcal{L}$ [see (6)].

Specifically, *MentorNet* employs an LSTM architecture to calculate the loss weight $\omega = \mathcal{G}(\mathcal{L}, \vec{z}, \theta)$ based on sample loss $\mathcal{L}$ and iteration variables $\vec{z}$, where θ denotes its parameters. The weighted loss $\mathcal{L}^M$ is computed as (7)

$$\mathcal{L}^M = \omega \mathcal{L}. \tag{7}$$

Therefore, the *MentorNet* subnet functions as a dynamic weighting module and utilizes a separate Adam optimizer to update its parameters.

Additionally, EEGNet without sample-reweighting serves as an independent baseline for comparison.

B. System Settings

The proposed system utilizes the following settings for prediction.

1) Evaluation Metrics: Classification *accuracies*, *F1-scores*, and *Matthews Correlation Coefficient (MCC-scores)* are calculated for each model to ensure the consistency between model outputs and given labels on clean data.

2) Validation Protocol: A *subject-independent* protocol with ten-fold cross-validation ensures comprehensive coverage of all 2190 samples, with mean evaluation metrics computed across the folds.

3) Dataset Splitting: As shown in Table I, for each validation fold, 12 samples (i.e. six from the TP set and six from the TN set) are randomly selected to form a *test set*[1] For further validation in Section VI-C, 3 TP samples are randomly selected from the TP set. Consequently, the remaining 21 TP and 15 TN samples form the *training set*. To enhance data diversity and address label imbalance, six UN samples with negative labels are also included. While most UN samples likely do not exhibit SI, they contribute to category representation. For UN samples potentially containing SI, employed antinoise techniques help mitigate their influence through weighted filtering.

[1]To ensure accurate evaluation metric values, samples with label uncertainty (UN) are excluded from the test set.

TABLE II
ACCURACIES OF FOUR MODELS ON THE PREDICTION TASK

	PPG	GSR	FUSE
Baseline [39]	0.5750 ± 0.0970	0.5167 ± 0.0311	0.6250 ± 0.1219
MentorNet [37]	0.6167 ± 0.1011	0.5833 ± 0.0865	0.7000 ± 0.1083
Co-teaching [38]	0.6333 ± 0.1243	0.6000 ± 0.0970	0.6917 ± 0.1145
DBN_ConvNet	**0.6833 ± 0.1185**	**0.7083 ± 0.1049**	**0.7667 ± 0.0865**
Average	0.6417	0.6083	0.6933

Note: The bold entries indicate the optimal value for each respective column.

TABLE III
F1 SCORES OF FOUR MODELS ON THE PREDICTION TASK

	PPG	GSR	FUSE
Baseline [39]	0.5895 ± 0.2979	0.2706 ± 0.3367	0.5540 ± 0.2860
MentorNet [37]	0.5342 ± 0.0875	0.6074 ± 0.1495	0.7120 ± 0.1680
Co-teaching [38]	*	*	*
DBN_ConvNet	**0.6215 ± 0.0868**	**0.6518 ± 0.1557**	**0.7435 ± 0.0997**
Average	0.5817	0.5099	0.6698

Note: * Indicates that the value is very close to 0. The bold entries indicate the optimal value for each respective column.

TABLE IV
MCC SCORES OF FOUR MODELS ON THE PREDICTION TASK

	PPG	GSR	FUSE
Baseline [39]	0.2231 ± 0.2133	0.0302 ± 0.1125	0.2918 ± 0.2614
MentorNet [37]	0.2970 ± 0.0275	0.2400 ± 0.0179	0.4424 ± 0.0152
Coteaching [38]	*	*	*
DBN_ConvNet	**0.4265 ± 0.1990**	**0.4520 ± 0.2169**	**0.5758 ± 0.1634**
Average	0.3155	0.2407	0.4367

Note: * Indicates that the value is very close to 0. The bold entries indicate the optimal value for each respective column.

4) Generating Prediction Results: In each fold of the cross-validation, the trained model is used to evaluate not only the samples in the test set but also all UN samples (approximately 2000) not included in the training set, to obtain their SI probability scores. The average score for each participant across all folds is calculated as their final prediction result.

VI. EXPERIMENT

This section summarizes the results obtained from three classification models employed in the prediction system (Section VI-A). Furthermore, a comparison of these results with state-of-the-art approaches is presented (Section VI-B), followed by an in-depth analysis of the key components of DBN_ConvNet through ablation studies (Section VI-C). We also examine misclassified cases and the underlying reasons for these errors further in Appendix G (available online).

A. Classification Results

In Tables II–IV, the average values and corresponding standard deviations of accuracy, F1 score, and MCC score are reported for four models: the three classification models presented in Section V-A and the EEGNet model [39] as the baseline. The proposed DBN_ConvNet achieves the best performance in all these three modalities (i.e., *PPG*, *GSR*, and *FUSE*). Specifically, in the *FUSE* modality, DBN_ConvNet achieves an accuracy of 76.67% (which is 14.17% higher than baseline), an F1 score of 74.35% (which is 18.95% higher than baseline), and an MCC score of 57.58% (which is 28.40% higher than baseline).

The integration of loss reweighting strategies into EEGNet markedly improves noise mitigation, with *MentorNet* [37] and *Coteaching* [38] increasing accuracy by 7.50% and 6.67%, respectively. Despite these gains, *Coteaching* records lower scores in Tables III to IV, likely due to overfitting caused by small mini-batch sizes used for sample exchange [74].

Furthermore, integrating GSR and PPG features (FUSE modality) enhances the classification performance for most of the evaluated models. Consequently, subsequent investigations mainly concentrate on exploring the potential of the FUSE modality.

B. Comparison With State-of-the-Arts

To demonstrate the efficacy of the proposed methods in handling noise, they are compared with three established machine learning algorithms, four predominant deep learning methods designed for physiological signals, and three state-of-the-art noisy learning architectures, all using the *FUSE* modality.

1) Machine Learning (ML) Methods: Support Vector Machines (SVM), *K-Nearest Neighbors* (KNN), and *Random Forest* (RF) from the Scikit-learn library [75] handle our 2-D time-series data, which is flattened into 1-D vectors. Each participant's data is segmented into 70 input vectors, each of dimensionality $D_{(73)}^{\mathrm{FUSE}}$, ensuring minimal information loss while preserving feature correlations.

2) Deep Learning (DL) Methods: FBCNet [59], *DeepConvNet* [57], *ShallowConvNet* [57], *DGCNN* [60], and LSTM [76] are selected as state-of-the-arts. The first four models, along with the baseline EEGNet model, utilize convolution layers and fully connected layers for the classification task.

3) Noisy Learning (NL) Methods: CTW [33], *SREA* [34], and *DivideMix* [35] are employed as recent advancements in handling noisy labels, utilizing techniques that segregate clean and noisy data or adaptively reweight training samples to improve robustness in learning from noisy datasets.

Implementation details of ML, DL, and NL methods are detailed in Appendix D (available online). A Mann–Whitney U Test is conducted to assess statistical significance between the results of DBN_ConvNet and of other methods, further detailed in Appendix F (available online).

The results, including accuracies, F1 scores, and MCC scores of these comparative classifiers, together with the p-values, are presented in Table V, clearly demonstrating the superiority of the proposed DBN_ConvNet model.

TABLE V
ACCURACIES, F1-SCORES, MCC-SCORES, AND p-VALUES OF COMPARATIVE METHODS

	Accuracy	F1	MCC	p
KNN	0.5690 ± 0.0709	0.6027 ± 0.0633	0.1388 ± 0.1438	0.0034**
RF	0.5833 ± 0.0978	0.6009 ± 0.0965	0.1706 ± 0.2008	0.0012**
SVM	0.5976 ± 0.0621	0.6441 ± 0.1632	0.2070 ± 0.1440	0.0001**
FBCNet [59]	0.6250 ± 0.0822	0.5761 ± 0.2909	0.2802 ± 0.2180	0.0001**
EEGNet [39]	0.6250 ± 0.1219	0.5540 ± 0.2860	0.2918 ± 0.2614	0.0001**
DeepConvNet [57]	0.5750 ± 0.0990	0.6091 ± 0.2822	0.1891 ± 0.2309	0.0054**
ShallowConvNet [57]	0.6250 ± 0.1219	0.6108 ± 0.1857	0.2988 ± 0.2572	0.0001**
DGCNN [60]	0.6083 ± 0.1039	0.5832 ± 0.2798	0.2463 ± 0.2166	0.0004**
LSTM	0.6000 ± 0.1233	0.4516 ± 0.1018	0.2410 ± 0.2553	0.0001**
CTW [33]	0.5076 ± 0.1134	0.5015 ± 0.1200	0.0174 ± 0.2450	0.0980
SREA [34]	0.5606 ± 0.1180	0.5525 ± 0.1211	0.1435 ± 0.2406	0.0020**
DivideMix [35]	0.5727 ± 0.1531	0.5473 ± 0.1893	0.1457 ± 0.3299	0.0660
DBN_ConvNet	**0.7667 ± 0.0865**	**0.7435 ± 0.0997**	**0.5758 ± 0.1634**	———

Note: * $p < 0.05$ ** $p < 0.01$. The bold entries indicate the optimal value for each respective column.

TABLE VI
ACCURACIES, F1-SCORES, AND MCC-SCORES OF DIFFERENT CONDITIONS FOR THE SELFSUPERVISED TRAINING PHASE

Condition	Accuracy	F1	MCC
Without DBN Pretraining	0.7251 ± 0.1025	0.6669 ± 0.1200	0.4830 ± 0.2064
With UN Samples	0.7500 ± 0.0795	0.7084 ± 0.1505	0.5336 ± 0.1507
With Entire Training Set	**0.7667 ± 0.0865**	**0.7435 ± 0.0997**	**0.5758 ± 0.1634**

Note: The bold entries indicate the optimal value for each respective column.

C. Ablation Study

To demonstrate the effectiveness of the design principles sustaining DBN_ConvNet, specifically the integration of self-supervised DBN training as outlined in Section IV-B1 and the implementation of confidence learning (CL) for noise filtering detailed in Section IV-B2, we conduct a comprehensive series of five ablation experiments.

1) Self-Supervised Training of DBN: The DBN_ConvNet model significantly outperforms other methods in the FUSE modality, as evident in Tables II and V. This superior performance may be attributed to the self-supervised training phase of the DBN module, which effectively reduces redundancy in the feature sequences while preserving essential components to minimize reconstruction loss.

To verify the analysis of the self-supervised training process above, two additional experiments are designed. First, the self-supervised phase is entirely removed, resulting in a unidirectional data flow through the DBN module, the convolution module, and the classification module of DBN_ConvNet in sequence, and the model is trained with full supervision. Second, only UN samples are used in the self-supervised phase to test the data purity's influence on the DBN module's reconstruction function. This experiment is conducted with the same validation protocols as the original process in Section IV-B2, and the results are summarized in Table VI, which shows that with

TABLE VII
ACCURACIES, F1-SCORES, AND MCC-SCORES IN CROSS-VALIDATION BEFORE AND AFTER CL METHOD

Stage	Accuracy	F1	MCC
Before CL	0.6833 ± 0.0900	0.6491 ± 0.1023	0.4169 ± 0.1872
After CL	**0.7667 ± 0.0865**	**0.7435 ± 0.0997**	**0.5758 ± 0.1634**

Note: The bold entries indicate the optimal value for each respective column.

TABLE VIII
RANKS OF THE 3-MIXED TP SAMPLES

No. Validation	Prisoner ID	Rank	Average	Overall Average
1	A	0	39.33	**41.33**
	B	17		
	C	101		
2	D	9	58.33	
	E	59		
	F	107		
3	G	0	26.33	
	H	28		
	I	51		

Note: The bold entries indicate the optimal value for each respective column.

TABLE IX
PERFORMANCE OF DBN_CONVNET WITH DIFFERENT WINDOW LENGTHS

	Accuracy	F1	MCC
CUT	0.7158 ± 0.0472	0.6975 ± 0.0716	0.4385 ± 0.0903
ALL	**0.7667 ± 0.0865**	**0.7435 ± 0.0997**	**0.5758 ± 0.1634**

Note: "ALL" stands for the original length 70, "CUT" for the sliced length 7. The bold entries indicate the optimal value for each respective column.

self-supervised training, DBN_ConvNet performs much better. When trained with only UN samples, the DBN module shows slightly inferior reconstruction ability compared to training with the entire training set but still outperforms the nonreconstruction condition. This suggests that the features extracted from *UN* samples are exploitable, although the decrease in data purity may introduce challenges for self-supervised learning via reconstruction.

2) Effectiveness of CL Method: The DBN_ConvNet training process consists of two stages, as outlined in Section IV-B, with the CL method serving as an intermediate noise-filtering step. To evaluate its impact, performances before and after applying the CL method are compared. Table VII demonstrates an improvement in classification results after discarding *UN* samples identified as noisy by the CL method, confirming the mismatch between their given labels and ground truth.

3) Validation of the Detected Noise: As detailed in Section V-B, 3 TP samples are excluded from both training and testing in each cross-validation fold. These samples, along with UN samples not used for training, are assessed by the trained models to predict SI probabilities. The effectiveness of noise detection is evaluated by analyzing whether these TP samples rank within the top 200 of approximately 2000 UN samples. This ratio, reflective of the typical prison suicide rate, suggests a manageable scope for further diagnosis. This validation process is repeated three times, each time with a different set of 3 TP samples, resulting in nine individuals being tested. The ranking outcomes, detailed in Table VIII, support the accomplishment of the task as defined in Section I.

4) Impact of Dataset Size: The relatively small number of TP and TN labels necessitated splitting feature sequences into smaller temporal segments to enlarge the test set. As shown in Table IX, dividing the original feature sequence of the FUSE modality into ten segments with DBN_ConvNet reduces the length from 70 to 7. Although this results in a slight performance drop, it remains satisfactory. The diminished metrics likely stem from the reduced capacity of the model to leverage complete temporal contexts, essential for capturing comprehensive dynamics within the data.

5) Validation on Public Datasets: To assess the generalizability of DBN_ConvNet, we test on three binary-classification EEG datasets from the UEA repository—EyesOpenShut, FingerMovements, and SelfRegulationSCP1 [77]. These datasets were chosen for their task relevance and similar sample sizes to our study. We introduce symmetrical artificial noise at a ratio of 0.3 to each dataset. For DBN_ConvNet, hyperparameters D^m and T are adjusted to fit the dimensionality and time length of these datasets. As shown in Table X, our DBN_ConvNet achieves the highest accuracy on all these three datasets, further underscoring its robustness and adaptability.

VII. DISCUSSION

In this section, we aim to provide a comprehensive analysis of the insights gained from our model, which has been instrumental in unraveling the intricate relationship between physiological indicators and the prediction of SI. Additionally, we discuss the limitations of our study and provide suggestions for future research.

A. Unveiling Potential Physiological Indicators: Insights From the Model

Analysis of DBN_ConvNet's parameters in the FUSE modality, which achieved the best prediction performance, reveals key physiological indicators of SI. Specifically, we calculate the average of absolute values of the convolutional weights from the first convolutional block of DBN_ConvNet [refer to Appendix C, Table A2 for detailed information (available online)] across the filters, resulting in a set of $D^m = 73$ values that represent the weight assigned to each feature. These values are then sorted in descending order and provided in Appendix E, Table A3 (available online). These statistical findings obtained through deep learning are consistent with previous empirical observations and support their relevance in understanding SI.

Significant heart rate variability (HRV) indicators from the PPG modality, including pNN20, S, SDNN, and SD1 [detailed in Appendix A, Table A1 (available online)], rank

TABLE X
ACCURACIES, F1-SCORES AND MCC SCORES OF COMPARATIVE METHODS ON NOISED PUBLIC SERIAL DATASETS

Model_Name	Eyes Open Shut			Finger Movements			Selfregulation SCP1		
	Accuracy	F1	MCC	Accuracy	F1	MCC	Accuracy	F1	MCC
FBCNet [59]	0.5286 ± 0.0797	0.5414 ± 0.2955	0.0706 ± 0.1677	0.5111 ± 0.0387	0.5934 ± 0.1588	0.0291 ± 0.0899	0.5148 ± 0.0406	0.3241 ± 0.3169	0.0445 ± 0.0922
EEGNet [39]	0.6643 ± 0.1249	0.6484 ± 0.1362	0.3341 ± 0.2510	0.5444 ± 0.0548	0.6271 ± 0.0829	0.1082 ± 0.1243	0.8389 ± 0.0284	0.8397 ± 0.0230	0.6813 ± 0.0552
DeepConvNet [57]	0.6786 ± 0.0684	0.6843 ± 0.1128	**0.4373 ± 0.1225**	0.5028 ± 0.0060	0.5459 ± 0.2690	0.0097 ± 0.0221	0.6074 ± 0.0450	0.6800 ± 0.0267	0.2505 ± 0.0806
ShallowConvNet [57]	0.6929 ± 0.1015	**0.7035 ± 0.1059**	0.4050 ± 0.2123	0.5264 ± 0.0266	0.6042 ± 0.0819	0.0657 ± 0.0733	0.6463 ± 0.0791	0.6323 ± 0.1051	0.3195 ± 0.1740
DGCNN [60]	0.6357 ± 0.1266	0.6108 ± 0.1842	0.2875 ± 0.2617	0.5625 ± 0.0441	0.6042 ± 0.0484	0.1298 ± 0.0931	0.8519 ± 0.0283	0.8540 ± 0.0244	0.7061 ± 0.0580
LSTM	0.5643 ± 0.0616	0.4566 ± 0.2461	0.1706 ± 0.1684	0.5375 ± 0.0571	0.5625 ± 0.2242	0.0972 ± 0.1236	0.5648 ± 0.1201	0.5627 ± 0.2851	0.1620 ± 0.2512
CTW [33]	0.5418 ± 0.1986	0.5366 ± 0.1982	0.0774 ± 0.4044	0.5092 ± 0.0275	0.5072 ± 0.0281	0.0188 ± 0.0534	0.7686 ± 0.0397	0.7678 ± 0.0408	0.5412 ± 0.0739
SREA [34]	0.3934 ± 0.1470	0.3708 ± 0.1506	0.0041 ± 0.4002	0.4538 ± 0.0235	0.4465 ± 0.0279	0.0045 ± 0.4435	0.7629 ± 0.0440	0.7627 ± 0.0440	0.0076 ± 0.7644
DivideMix [35]	0.3769 ± 0.0985	0.3747 ± 0.0991	-0.2544 ± 0.2050	0.4652 ± 0.0507	0.4506 ± 0.0404	-0.0713 ± 0.1019	0.6885 ± 0.0557	0.6647 ± 0.0906	0.4092 ± 0.0756
Co-teaching [38]	0.6500 ± 0.0542	0.5157 ± 0.0910	0.1584 ± 0.0822	0.5653 ± 0.0413	0.6124 ± 0.0495	0.0440 ± 0.0714	0.8519 ± 0.0428	0.6907 ± 0.0338	0.3921 ± 0.0664
Mentor [37]	0.6286 ± 0.1160	0.6166 ± 0.1418	0.2624 ± 0.2326	0.5542 ± 0.0434	**0.6499 ± 0.0324**	0.1327 ± 0.1014	0.8148 ± 0.0600	0.8089 ± 0.0577	0.6359 ± 0.1202
DBN_ConvNet (ours)	**0.7071 ± 0.0943**	0.6819 ± 0.1144	0.4321 ± 0.1985	**0.6097 ± 0.0608**	0.6476 ± 0.0627	**0.2284 ± 0.1244**	**0.8537 ± 0.0550**	**0.8563 ± 0.0544**	**0.7105 ± 0.1097**
MeanBaselines	0.5777 ± 0.1070	0.5509 ± 0.1601	0.1775 ± 0.2280	0.5211 ± 0.0376	0.5640 ± 0.0949	0.0517 ± 0.1180	0.7192 ± 0.0531	0.6898 ± 0.0953	0.3773 ± 0.1647
Mean	0.5885 ± 0.1059	0.5618 ± 0.1563	0.1988 ± 0.2255	0.5285 ± 0.0395	0.5710 ± 0.0922	0.0664 ± 0.1185	0.7304 ± 0.0532	0.7037 ± 0.0919	0.4050 ± 0.1601

Note: "Mean-Baselines" represents the average performance of all models excluding DBN_ConvNet, while "Mean" reflects the overall average including DBN_ConvNet. The "Average" column presents each model's average performance across the four datasets. The bold entries indicate the optimal value for each respective column.

among the top 30 discriminative features. Previous research has highlighted that individuals with SI or historical suicide attempts exhibit significantly lower HRV measures such as SDNN and RMSSD, both during resting states and emotional exposures, underscoring their potential as reliable markers for SI compared to control groups [26], [78], [79], [80], [81].

In the GSR modality, key features such as the maximum, median, and standard deviation of skin conductance response (SCR) rank among the top 20 indicators. SCR, reflecting the phasic activity of GSR, is sensitive to rapid changes triggered by stimuli. Research indicates that individuals with depression or suicidal behavior exhibit faster SCR habituation to acoustic stimuli, suggesting SCR's potential as a marker for acute suicide risk [82].

Moreover, deep-learning models have identified additional crucial features, like GSR derivative and PPG signal amplitude. The ability of neural networks to detect these subtle fluctuations enhances our understanding of the physiological manifestations associated with the emergence and development of SI, shedding light on the complex interplay between physiological responses and mental health.

B. Limitations and Future Work

While our current research focuses on physiological signals for predicting suicidal ideation, future studies will employ a multimodal framework incorporating demographic data, criminal histories, and offense categories, which are significant risk factors in correctional environments [83]. To enhance the generalizability and robustness of our model, we plan to adapt our methods across diverse age and gender groups for improved label accuracy. Furthermore, we aim to incorporate comprehensive datasets such as the Adolescent Brain Cognitive Development® (ABCD, https://abcdstudy.org), enriched with extensive demographic and behavioral data. Additionally, we will explore using EEG and other physiological data modalities to address hardware variability and enhance data precision.

VIII. CONCLUSION

In this study, we have developed models robust against label noise for predicting suicidal ideation. Our novel DBN_ConvNet architecture, integrating confidence learning [36], is designed to handle physiological data effectively. Additionally, we enhanced the widely adopted EEGNet [39] by incorporating noise-resilient reweighting strategies, namely MentorNet [37] and Coteaching [38], thereby developing a dual solution system for SI prediction.

Experimental results confirm that our approaches significantly improve prediction accuracy and F1 score compared to traditional machine learning methods, state-of-the-art deep learning, and noisy learning models. The DBN_ConvNet architecture, in particular, demonstrates superior performance by leveraging the integrated analysis of GSR and PPG signals, thus emerging as the most effective model in our evaluations.

DATA AVAILABILITY STATEMENT

This dataset is publicly available upon request from our website at https://github.com/YongjinLiu/Biometric-Suicidal-Ideation-Database.

REFERENCES

[1] M. A. Oquendo et al., "Occurrence and characteristics of suicidal ideation in psychiatrically healthy individuals based on ecological momentary assessment," *Mol. Psychiatry*, pp. 1–8, 2024.
[2] D. M. Corey and R. Borum, "Forensic assessment for high-risk occupations," in *Handbook of Psychology*, 2nded., vol. 11, 2012.
[3] A. Burton Denmark, E. Hess, and M. S. Becker, "College students' reasons for concealing suicidal ideation," *J. College Student Psychotherapy*, vol. 26, no. 2, pp. 83–98, 2012.
[4] A. C. Wu, D. Donnelly-McLay, M. G. Weisskopf, E. McNeely, T. S. Betancourt, and J. G. Allen, "Airplane pilot mental health and suicidal thoughts: a cross-sectional descriptive study via anonymous web-based survey," *Environmental Health*, vol. 15, no. 1, pp. 1–12, 2016.
[5] A. Liebling and H. Arnold, "Social relationships between prisoners in a maximum security prison: Violence, faith, and the declining nature of trust," *J. Criminal Justice*, vol. 40, no. 5, pp. 413–424, 2012.

[6] M. M. Nielsen, "Pains and possibilities in prison: On the use of emotions and positioning in ethnographic research," *Acta Sociol.*, vol. 53, no. 4, pp. 307–321, 2010.

[7] E. Laksana, T. Baltrušaitis, L.-P. Morency, and J. P. Pestian, "Investigating facial behavior indicators of suicidal ideation," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*. Piscataway, NJ, USA: IEEE Press, 2017, pp. 770–777.

[8] L. C. Barry, E. Coman, D. Wakefield, R. L. Trestman, Y. Conwell, and D. C. Steffens, "Functional disability, depression, and suicidal ideation in older prisoners," *J. Affect. Disorders*, vol. 266, pp. 366–373, 2020.

[9] L. C. Barry, D. B. Wakefield, R. L. Trestman, and Y. Conwell, "Active and passive suicidal ideation in older prisoners." *Crisis J. Crisis Intervention Suicide Prevention*, vol. 37, no. 2, p. 88, 2016.

[10] C. Figueroa Saavedra, T. Otzen Hernández, C. Alarcón Godoy, A. Ríos Pérez, D. Frugone Salinas, and R. Lagos Hernández, "Association between suicidal ideation and acoustic parameters of university students' voice and speech: A pilot study," *Logopedics Phoniatrics Vocology*, vol. 46, no. 2, pp. 55–62, 2021.

[11] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," *IEEE Trans. Comput. Social Syst.*, vol. 8, no. 1, pp. 214–226, Jan. 2020.

[12] X. Xu, "Detecting suicide ideation in the online environment: A survey of methods and challenges," *IEEE Trans. Comput. Social Syst.*, vol. 9, no. 3, pp. 679–687, Mar. 2021.

[13] T. D. Gunter, J. T. Chibnall, S. K. Antoniak, R. A. Philibert, and N. Hollenbeck, "Predictors of suicidal ideation, suicide attempts, and self-harm without lethal intent in a community corrections sample," *J. Criminal Justice*, vol. 39, no. 3, pp. 238–245, 2011.

[14] C. R. Pennington, R. J. Cramer, H. A. Miller, and J. S. Anastasi, "Psychopathy, depression, and anxiety as predictors of suicidal ideation in offenders," *Death Studies*, vol. 39, no. 5, pp. 288–295, 2015.

[15] F. J. Richie, J. Bonner, A. Wittenborn, L. M. Weinstock, C. Zlotnick, and J. E. Johnson, "Social support and suicidal ideation among prisoners with major depressive disorder," *Archives Suicide Res.*, vol. 25, no. 1, pp. 107–114, 2021.

[16] S. J. Garlow et al., "Depression, desperation, and suicidal ideation in college students: Results from the American foundation for suicide prevention college screening project at Emory university," *Depression Anxiety*, vol. 25, no. 6, pp. 482–488, 2008.

[17] L. K. McCorry, "Physiology of the autonomic nervous system," *Amer. J. Pharmaceutical Educ.*, vol. 71, no. 4, 2007.

[18] S. D. Kreibig, F. H. Wilhelm, W. T. Roth, and J. J. Gross, "Cardiovascular, electrodermal, and respiratory response patterns to fear-and sadness-inducing films," *Psychophysiology*, vol. 44, no. 5, pp. 787–806, 2007.

[19] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biol. Psychol.*, vol. 84, no. 3, pp. 394–421, 2010.

[20] T. D. Pham, T. C. Thang, M. Oyama-Higa, and M. Sugiyama, "Mental-disorder detection using chaos and nonlinear dynamical analysis of photoplethysmographic signals," *Chaos, Solitons Fractals*, vol. 51, pp. 64–74, 2013.

[21] A. Clamor, M. M. Hartmann, U. Koether, C. Otte, S. Moritz, and T. M. Lincoln, "Altered autonomic arousal in psychosis: An analysis of vulnerability and specificity," *Schizophrenia Res.*, vol. 154, no. 1–3, pp. 73–78, 2014.

[22] J. W. Choi, H. Thakur, and J. R. Cohen, "Cardiac autonomic functioning across stress and reward: Links with depression in emerging adults," *Int. J. Psychophysiology*, vol. 168, pp. 1–8, 2021.

[23] P. Grossman and E. W. Taylor, "Toward understanding respiratory sinus arrhythmia: Relations to cardiac vagal tone, evolution and biobehavioral functions," *Biological Psychology*, vol. 74, no. 2, pp. 263–285, 2007.

[24] J. Rottenberg, "Mood and emotion in major depression," *Curr. Directions Psychol. Sci.*, vol. 14, no. 3, pp. 167–170, 2005.

[25] J. Rottenberg, K. Salomon, J. J. Gross, and I. H. Gotlib, "Vagal withdrawal to a sad film predicts subsequent recovery from depression," *Psychophysiology*, vol. 42, no. 3, pp. 277–281, 2005.

[26] D. Adolph, T. Teismann, T. Forkmann, A. Wannemüller, and J. Margraf, "High frequency heart rate variability: Evidence for a transdiagnostic association with suicide ideation," *Biol. Psychol.*, vol. 138, pp. 165–171, 2018.

[27] L. Favril, F. Vander Laenen, C. Vandeviver, and K. Audenaert, "Suicidal ideation while incarcerated: Prevalence and correlates in a large sample of male prisoners in flanders, Belgium," *Int. J. Law Psychiatry*, vol. 55, pp. 19–28, 2017.

[28] S. Larney, L. Topp, D. Indig, C. O'driscoll, and D. Greenberg, "A cross-sectional survey of prevalence and correlates of suicidal ideation and suicide attempts among prisoners in New South Wales, Australia," *BMC Public Health*, vol. 12, no. 1, pp. 1–7, 2012.

[29] J. Zhang, V. E. Grabiner, Y. Zhou, and N. Li, "Suicidal ideation and its correlates in prisoners: A comparative study in China." *Crisis J. Crisis Intervention Suicide Prevention*, vol. 31, no. 6, p. 335, 2010.

[30] S. Fazel, R. Benning, and J. Danesh, "Suicides in male prisoners in England and Wales, 1978–2003," *Lancet*, vol. 366, no. 9493, pp. 1301–1302, 2005.

[31] A. Opitz-Welke, K. Bennefeld-Kersten, N. Konrad, and J. Welke, "Prison suicides in Germany from 2000 to 2011," *Int. J. Law Psychiatry*, vol. 36, no. 5–6, pp. 386–389, 2013.

[32] S. Fazel, T. Ramesh, and K. Hawton, "Suicide in prisons: An international study of prevalence and contributory factors," *Lancet Psychiatry*, vol. 4, no. 12, pp. 946–952, 2017.

[33] P. Ma, Z. Liu, J. Zheng, L. Wang, and Q. Ma, "CTW: Confident time-warping for time-series label-noise learning," in *Proc. 32nd Int. Joint Conf. Artif. Intell.*, 2023, pp. 4046–4054.

[34] A. Castellani, S. Schmitt, and B. Hammer, "Estimating the electrical power output of industrial devices with end-to-end time-series classification in the presence of label noise," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Springer, 2021, pp. 469–484.

[35] J. Li, R. Socher, and S. C. Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," 2020, *arXiv:2002.07394*.

[36] C. Northcutt, L. Jiang, and I. Chuang, "Confident learning: Estimating uncertainty in dataset labels," *J. Artif. Intell. Res.*, vol. 70, pp. 1373–1411, 2021.

[37] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, "Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *Int. Conf. Mach. Learn.* PMLR, 2018, pp. 2304–2313.

[38] B. Han et al., "Co-teaching: Robust training of deep neural networks with extremely noisy labels," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[39] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, 2016, Art. no. 056013.

[40] M. Chesin, M. Cascardi, W. Tsang, and S. Smith, "Blunted arousal in response to psychological stress is associated with current suicide ideation," *Arch. Suicide Res.*, vol. 24, no. 2, pp. S381–S390, 2020.

[41] S. Snitker, I. Macdonald, E. Ravussin, and A. Astrup, "The sympathetic nervous system and obesity: Role in aetiology and treatment," *Obesity Rev.*, vol. 1, no. 1, pp. 5–15, 2000.

[42] L. C. Mayes, "A developmental perspective on the regulation of arousal states," in *Seminars in Perinatology*, vol. 24, no. 4. Elsevier, 2000, pp. 267–279.

[43] M. Sharma, S. Kacker, and M. Sharma, "A brief introduction and review on galvanic skin response," *Int. J. Med. Res. Professionals*, vol. 2, no. 6, pp. 13–17, 2016.

[44] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiol. Meas.*, vol. 28, no. 3, p. R1, 2007.

[45] R. C. Brindle, A. C. Whittaker, A. Bibbey, D. Carroll, and A. T. Ginty, "Exploring the possible mechanisms of blunted cardiac reactivity to acute psychological stress," *Int. J. Psychophysiol.*, vol. 113, pp. 1–7, 2017.

[46] A. Schwerdtfeger and A.-K. Rosenkaimer, "Depressive symptoms and attenuated physiological reactivity to laboratory stressors," *Biol. Psychol.*, vol. 87, no. 3, pp. 430–438, 2011.

[47] K. Salomon, L. M. Bylsma, K. E. White, V. Panaite, and J. Rottenberg, "Is blunted cardiovascular reactivity in depression mood-state dependent? A comparison of major depressive disorder remitted depression and healthy controls," *Int. J. Psychophysiol.*, vol. 90, no. 1, pp. 50–57, 2013.

[48] D. Schneider et al., "Empathic behavioral and physiological responses to dynamic stimuli in depression," *Psychiatry Res.*, vol. 200, no. 2–3, pp. 294–305, 2012.

[49] L. M. Bylsma, "Emotion context insensitivity in depression: Toward an integrated and contextualized approach," *Psychophysiology*, vol. 58, no. 2, 2021, Art. no. e13715.

[50] A. Arsalan, M. Majid, S. M. Anwar, and U. Bagci, "Classification of perceived human stress using physiological signals," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*. Piscataway, NJ, USA: IEEE Press, 2019, pp. 1247–1250.

[51] X. Ding, X. Yue, R. Zheng, C. Bi, D. Li, and G. Yao, "Classifying major depression patients and healthy controls using EEG, eye tracking and galvanic skin response data," *J. Affect. Disorders*, vol. 251, pp. 156–161, 2019.

[52] J. A. Domínguez-Jiménez, K. C. Campo-Landines, J. C. Martínez-Santos, E. J. Delahoz, and S. H. Contreras-Ortiz, "A machine learning model for emotion recognition from physiological signals," *Biomed. Signal Process. Control*, vol. 55, 2020, Art. no. 101646.

[53] J. Lee, H. Lee, and M. Shin, "Driving stress detection using multimodal convolutional neural networks with nonlinear representation of short-term physiological signals," *Sensors*, vol. 21, no. 7, 2021, Art. no. 2381.

[54] S. Tiwari and S. Agarwal, "A shrewd artificial neural network-based hybrid model for pervasive stress detection of students using galvanic skin response and electrocardiogram signals," *Big Data*, vol. 9, no. 6, pp. 427–442, 2021.

[55] H. Zhu, G. Han, L. Shu, and H. Zhao, "Arvanet: Deep recurrent architecture for PPG-based negative mental-state monitoring," *IEEE Trans. Comput. Social Syst.*, vol. 8, no. 1, pp. 179–190, 2020.

[56] A. I. Siam, M. A. El-Affendi, A. Abou Elazm, G. M. El-Banby, N. A. El-Bahnasawy, F. E. Abd El-Samie, and A. A. Abd El-Latif, "Portable and real-time IOT-based healthcare monitoring system for daily medical applications," *IEEE Trans. Comput. Soc. Syst.*, 2022.

[57] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[58] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b," *Front. Neurosci.*, vol. 6, p. 39, 2012.

[59] R. Mane, N. Robinson, A. P. Vinod, S.-W. Lee, and C. Guan, "A multi-view CNN with novel variance layer for motor imagery brain computer interface," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*. Piscataway, NJ, USA: IEEE Press, 2020, pp. 2950–2953.

[60] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 532–541, 2018.

[61] G. Zhang, M. Yu, Y.-J. Liu, G. Zhao, D. Zhang, and W. Zheng, "SparseDGCNN: Recognizing emotion from multichannel EEG signals," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 537–548, Jan./Mar. 2021.

[62] A. Ghosh, H. Kumar, and P. S. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, 2017.

[63] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Adv. Neur. Inf. Process. Syst.*, vol. 31, 2018.

[64] H. Cheng, Z. Zhu, X. Li, Y. Gong, X. Sun, and Y. Liu, "Learning with instance-dependent label noise: A sample sieve approach," 2020, *arXiv:2010.02347*.

[65] G. Pleiss, T. Zhang, E. Elenberg, and K. Q. Weinberger, "Identifying mislabeled data using the area under the margin ranking," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 17044–17056, 2020.

[66] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.

[67] M. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," *Proc. 23rd Int. Conf. Neural Inf. Process.*, vol. 23, 2010.

[68] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, Cambridge, MA, USA: MIT Press, 1987, pp.194–281.

[69] G. Rao, R. A. Manaf, and H. S. Minhat, "A review on determinants of depression among adult prisoner," *Malaysian J. Med. Health Sciences*, vol. 14, no. 2, pp. 75–87, 2015.

[70] D. Pratt and E. Foster, "Feeling hopeful: Can hope and social support protect prisoners from suicide ideation?" *J. Forensic Psychiatry Psychol.*, vol. 31, no. 2, pp. 311–330, 2020.

[71] D. Makowski et al., "NeuroKit2: A python toolbox for neurophysiological signal processing," *Behav. Res. Methods*, vol. 53, no. 4, pp. 1689–1696, Feb. 2021. [Online]. Available: https://doi.org/10.3758%2Fs13428-020-01516-y

[72] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "cvxEDA: A convex optimization approach to electrodermal activity processing," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 4, pp. 797–804, Apr. 2015.

[73] V. G. Paul, d. B. Jonathan, J. Kris, and F. Glenn, "Paulvangentcom/heartrate_analysis_python 1.2.5." [Online]. Available: https://zenodo.org/record/3563546#.ZAdVH8dBy3A

[74] F. Yang et al., "Asymmetric co-teaching for unsupervised cross-domain person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 12597–12604.

[75] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[76] A. Graves, "Supervised sequence labelling with recurrent neural networks," in *Stud. Comput. Intell.*, 2012.

[77] A. Bagnall et al., "The UEA Multivariate Time Series Classification Archives," 2018, *arXiv:1811.00075*.

[78] C.-C. Chang, N.-S. Tzeng, Y.-C. Kao, C.-B. Yeh, and H.-A. Chang, "The relationships of current suicidal ideation with inflammatory markers and heart rate variability in unmedicated patients with major depressive disorder," *Psychiatry Res.*, vol. 258, pp. 449–456, 2017.

[79] A. H. Khandoker et al., "Predicting depressed patients with suicidal ideation from ECG recordings," *Med. Biol. Eng. Comput.*, vol. 55, pp. 793–805, 2017.

[80] A. Tsypes, K. M. James, M. L. Woody, C. Feurer, A. Y. Kudinova, and B. E. Gibb, "Resting respiratory sinus arrhythmia in suicide attempters," *Psychophysiology*, vol. 55, no. 2, 2018, Art. no. e12978. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/psyp.12978

[81] S. T. Wilson et al., "Heart rate variability and suicidal behavior," *Psychiatry Res.*, vol. 240, pp. 241–247, 2016.

[82] R. Calati, C. B. Nemeroff, J. Lopez-Castroman, L. J. Cohen, and I. Galynker, "Candidate biomarkers of suicide crisis syndrome: What to test next? A concept paper," *Int. J. Neuropsychopharmacology*, vol. 23, no. 3, pp. 192–205, 2020.

[83] S. Zhong et al., "Risk factors for suicide in prisons: A systematic review and meta-analysis," *Lancet Public Health*, vol. 6, no. 3, pp. e164–e174, 2021.