# Disentangled Adversarial Transfer Learning for Physiological Biosignals

## EMBC 2020

Mo Han[1], Ozan Ozdenizci[1], Ye Wang[2], Toshiaki Koike-Akino[2] and Deniz Erdogmus[1]

[1] Cognitive Systems Lab (CSL) - Northeastern University, Boston

[2] Mitsubishi Electric Research Laboratories (MERL), Cambridge

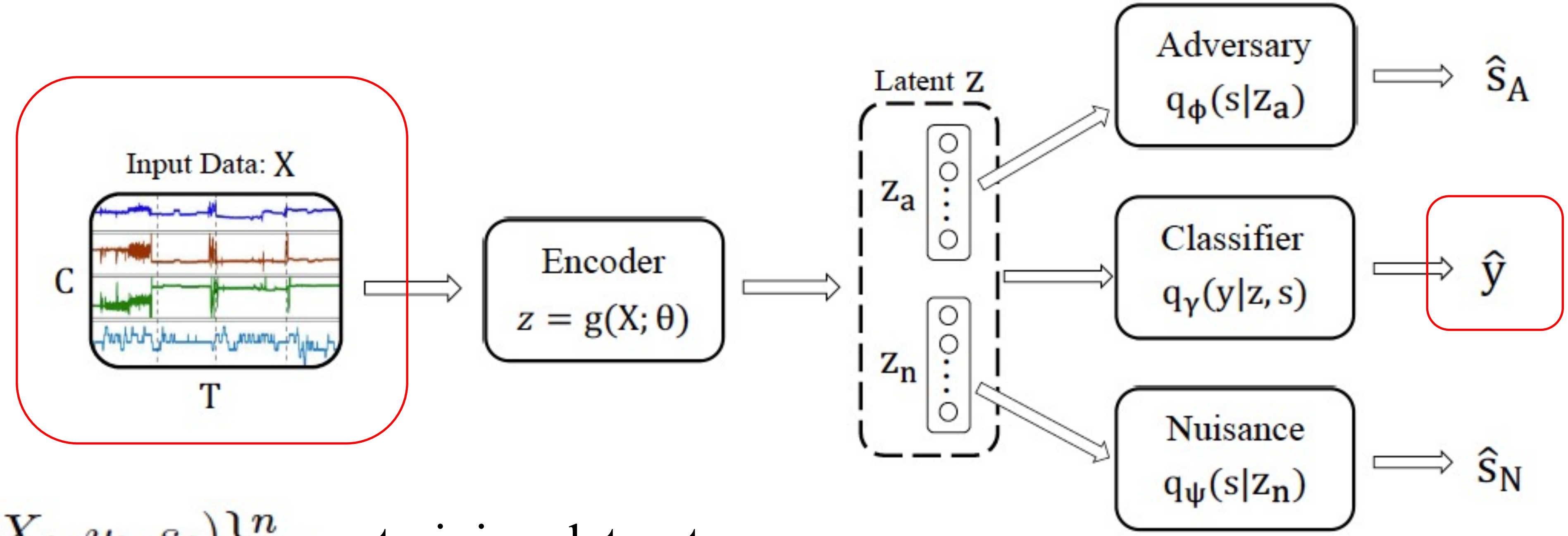
# 1. Introduction

- Physiological and Mental Status Monitoring
  - Traditional method: electroencephalography (EEG) signal
    - surface (non-invasive) or implanted (invasive) electrodes
    - frequent calibration
  - Non-EEG physiological biosignals: temperature, heart rate, and arterial oxygen, etc.
    - wrist-worn platform
    - more effective, comfortable, and less expensive
  - Major issue: variability among different subjects or recording sessions
- Transfer Learning
  - Cope with the change in data distributions, in order to fit a wider range of users
  - Adversarial training
    - allow the representation to predict dependent variables
    - simultaneously taking advantage of an adaptive measure
    - control the extent of its dependency during training

# 1. Introduction

- Our work: adversarial inference approach
  - Exploit disentangled nuisance-robust representations
  - Trade-off between task-related features and person-discriminative information
  - Additional censoring network blocks: **Adversary block and Nuisance block**
    - jointly train the adversary, nuisance and classifier units
    - task-discriminative features are incorporated for unknow users dissimilar from training data
    - features from known subjects are projected to unknow but similar users' data
  - Proposed disentangled adversarial transfer learning is applicable to other deep learning network approaches that are available

# 2. Methods: Disentangled Adversarial Transfer Learning



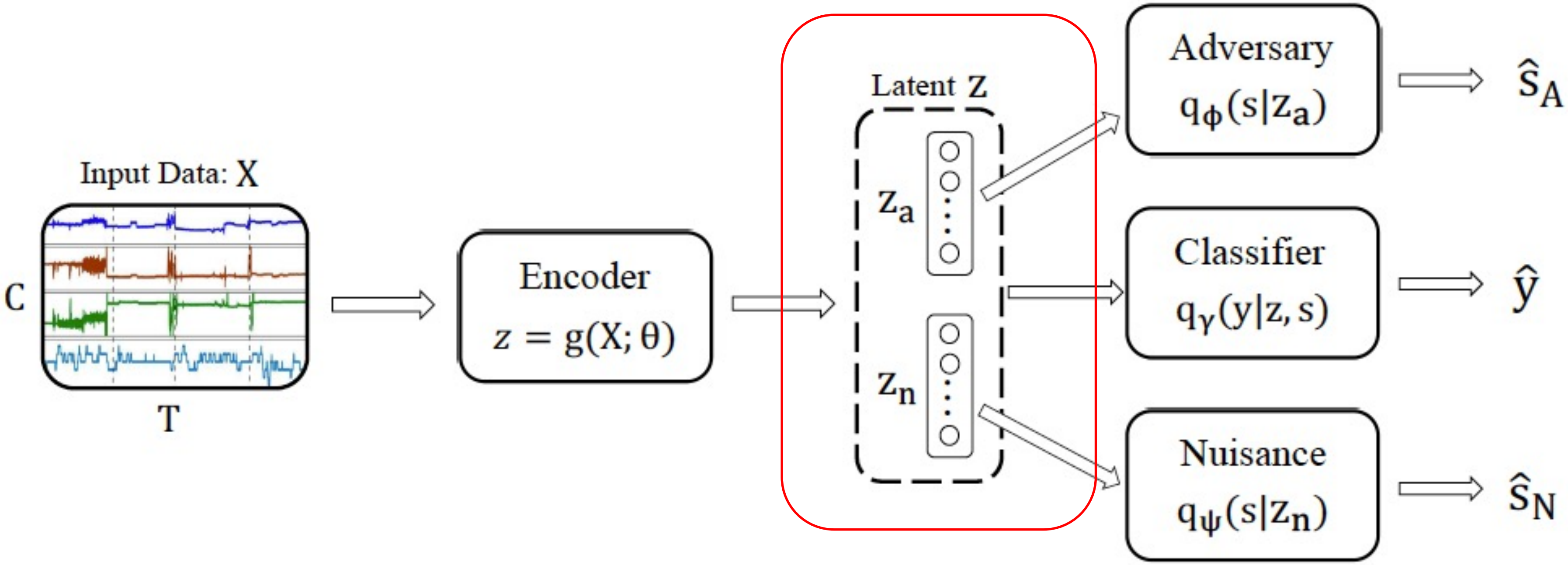

$\{(X_i, y_i, s_i)\}_{i=1}^{n}$ : training dataset

$X_i \in \mathbb{R}^{C\times T}$ : raw data at trial *i* recorded from *C* dimensions for *T* time samples

$y_i \in \{0, 1, \ldots, L-1\}$: label of user stress level status or task among *L* categories

$s_i \in \{1, \ldots, S-1, S\}$: subject identification (ID) among S individuals

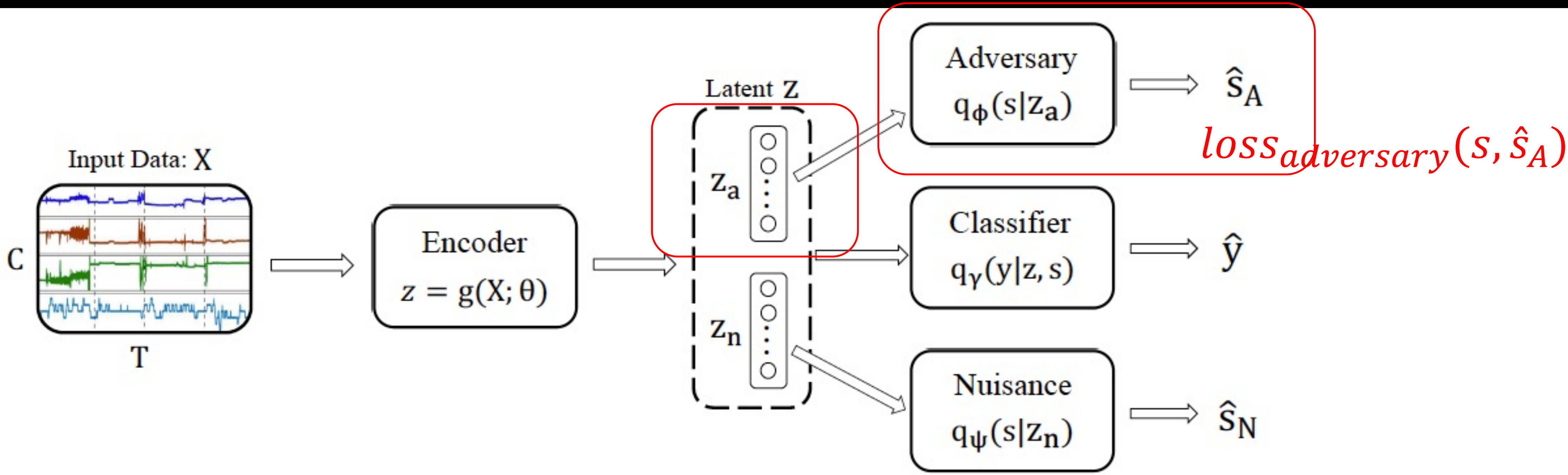
# 2. Methods: Disentangled Adversarial Transfer Learning




$z = g(X; \theta)$ : encoder, to learn the latent representation $z$ from data $X$

$z$ : latent feature, concatenation of $z_a$ and $z_n$ on a ratio of $(1\text{-}r_N) : r_N$

# 2. Methods: Disentangled Adversarial Transfer Learning

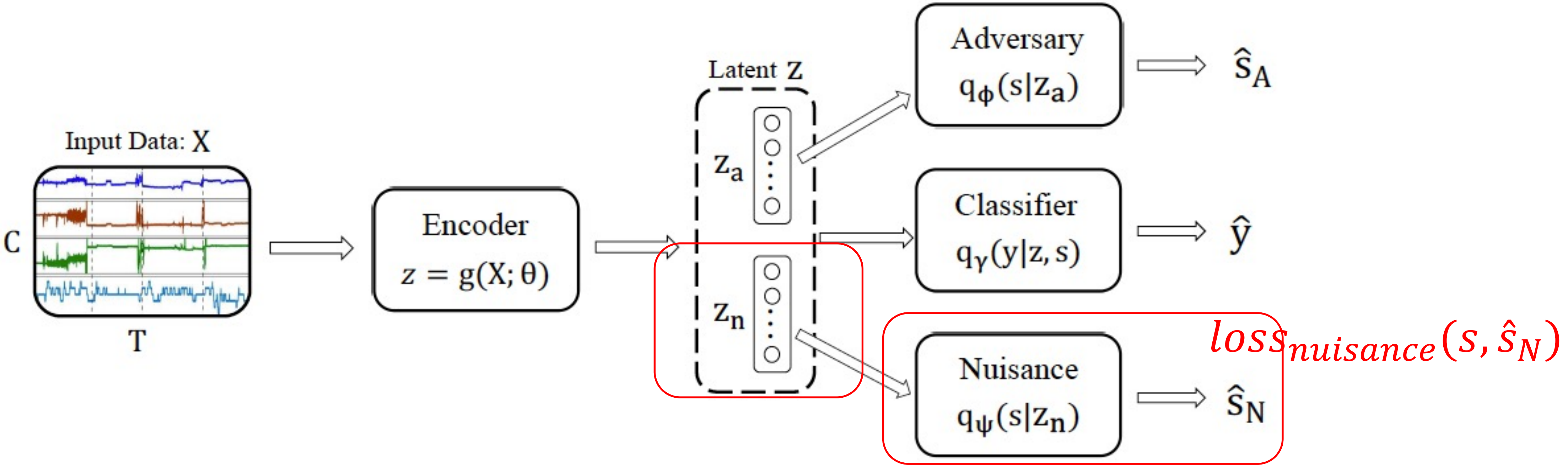



$z_a$ : input to the ***adversary*** network, aims to **conceal user-related information *s***

Adversary : a classifier for user-related information *s*, with $\hat{s}_A$ as the output

⇨ let feature $z_a$ have a lower correlation on classifying $s$, i.e. **maximize** $loss_{adversary}(s, \hat{s}_A)$

# 2. Methods: Disentangled Adversarial Transfer Learning

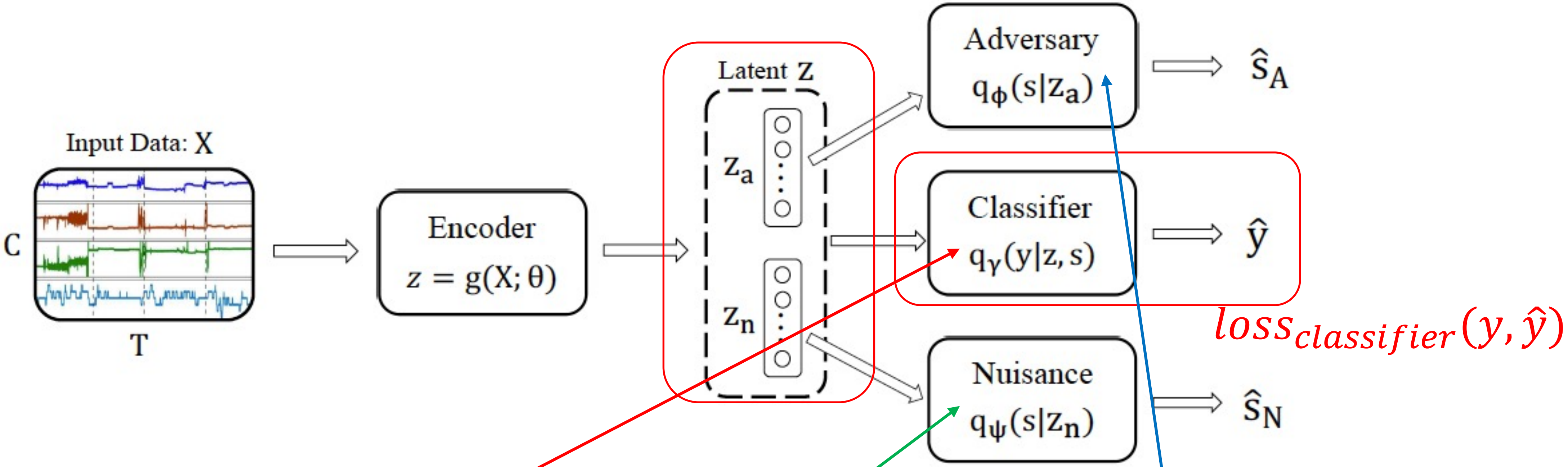



$z_n$ : input to the ***nuisance*** network, aims to **include user-related information *s***

Nuisance : a classifier for user-related information *s*, with $\hat{s}_N$ as the output

⇨ let feature $z_n$ have a higher correlation on classifying $s$, i.e. **minimize** $loss_{nuisance}(s, \hat{s}_N)$

# 2. Methods: Disentangled Adversarial Transfer Learning




$$\max_{\theta,\gamma,\psi} \min_{\phi} \quad \mathbb{E}\left[\log q_\gamma\left(y|g(X;\theta), s\right) + \lambda_N \log q_\psi\left(s|z_n\right) - \lambda_A \log q_\phi\left(s|z_a\right)\right]$$

min $loss_{classifier}(y, \hat{y})$ min $loss_{nuisance}(s, \hat{s}_N)$ max $loss_{adversary}(s, \hat{s}_A)$

# 3. Experimental Evaluation and Results: Physiological Biosignal Dataset

- Dataset: physiological biosignal dataset for assessing human stress status levels
  - 4 stress status (L = 4):
    (i). physical stress (ii). cognitive stress (iii). emotional stress (iv). relaxation
  - 20 healthy subjects (S = 20)
  - 7 channels (C = 7): biosensors containing
    (i). electrodermal activity (ii). temperature (iii). heart rate (iv). arterial oxygen, (v-vii). acceleration
  - 300 time samples (T = 300): task of 5 minutes downsampled to 1 Hz

# 3. Experimental Evaluation and Results: Experiment Implementation

- Parameters:
  - known: channel number C = 7, time sample T = 300, label number L = 4, subject number S = 20
  - to be optimized: adversary regularization weight $\lambda_A$ and nuisance regularization weights $\lambda_N$
  - to be optimized: nuisance representation rate $r_N$ among all features
- Parameter optimization:
  - 1. first optimize $\lambda_A$ with only adversary block: $\lambda_A \in \{0.05, 0.1\}$ with $\lambda_N = 0$ and $r_N = 0$
  - 2. fix the nuisance rate to $r_N = 0.2$: assume that the subject-related feature $z_N$ accounts for a small proportion among feature $z$ and keeps constant for all users and tasks
  - 3. second optimize $\lambda_N$ with both adversary and nuisance blocks: $\lambda_N \in \{0.001, 0.005, 0.05, 0.01, 0.2\}$ with $r_N = 0.2$ and optimized $\lambda_A$ from step 1
- Validation: cross-subjects validation using a leave-one-subject-out approach

# 3. Experimental Evaluation and Results: Results and Discussion

| | $\lambda_A$ | $\lambda_N$ | $r_N$ | Main Classifier | Adversary Network | Nuisance Network |
|---|---|---|---|---|---|---|
| Non-Adversarial | 0 | 0 | 0 | 79.88% | 71.13% | 6.17% |
| Adversarial | 0.005 | 0 | 0 | 79.97% | 35.62% | 6.15% |
| | 0.1 | 0 | 0 | 80.34% | 8.08% | 6.20% |
| **Disentangled Adversarial** | 0.1 | 0.001 | 0.2 | 80.62% | 7.05% | 39.03% |
| | **0.1** | **0.005** | **0.2** | **80.66%** | **7.90%** | **55.54%** |
| | 0.1 | 0.05 | 0.2 | 80.04% | 7.37% | 78.83% |
| | 0.1 | 0.1 | 0.2 | 80.36% | 8.08% | 83.72% |
| | 0.1 | 0.2 | 0.2 | 80.22% | 8.05% | 87.26% |

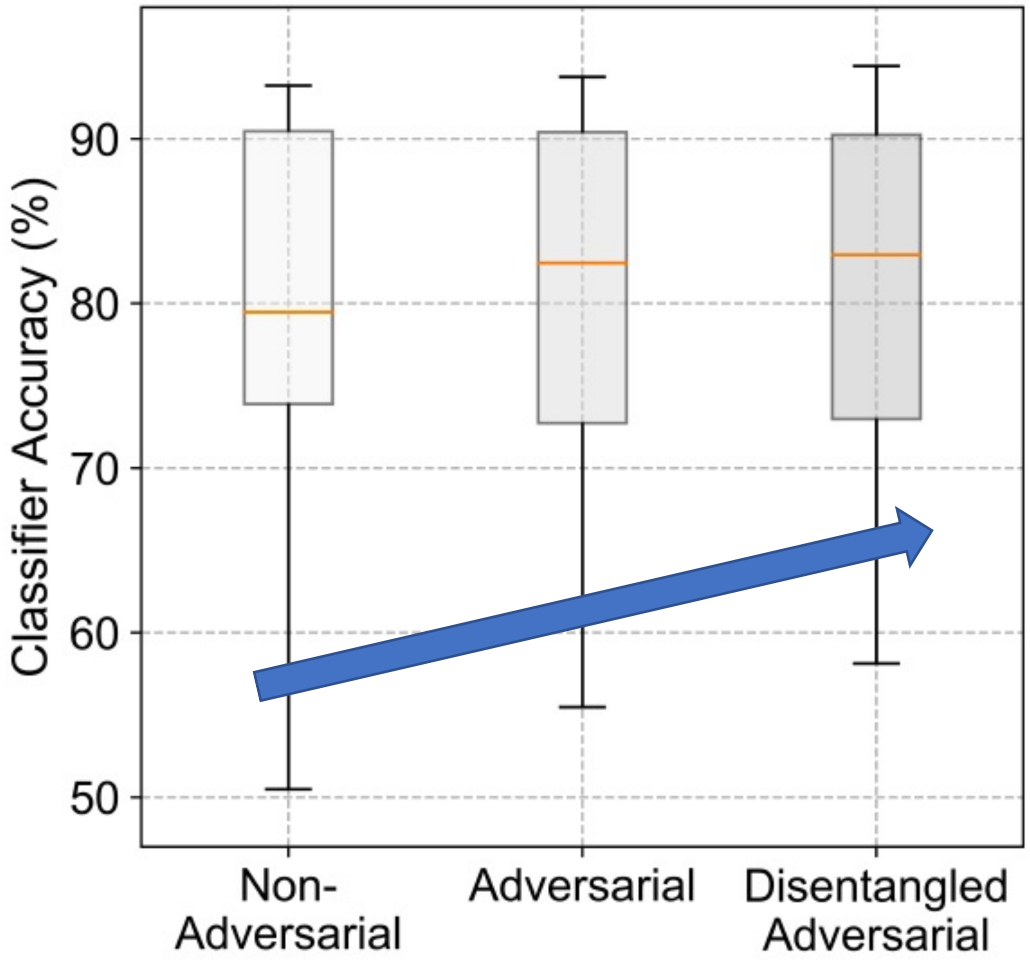
- Main classifier accuracy: 4-class decoding of human stress
  - preferable: higher, indicates better discrimination of stress status levels
- Adversary network accuracy: 20-class decoding of subject ID
  - preferable: lower, indicates less subject-specific information are preserved in feature $z_a$
- Nuisance network accuracy: 20-class decoding of subject ID
  - preferable: higher, indicates more subject-specific information are preserved in feature $z_n$

# 3. Experimental Evaluation and Results: Results and Discussion

| | $\lambda_A$ | $\lambda_N$ | $r_N$ | Main Classifier | Adversary Network | Nuisance Network |
|---|---|---|---|---|---|---|
| Non-Adversarial | 0 | 0 | 0 | 79.88% | 71.13% | 6.17% |
| Adversarial | 0.005 | 0 | 0 | 79.97% | 35.62% | 6.15% |
| | 0.1 | 0 | 0 | 80.34% | 8.08% | 6.20% |
| **Disentangled Adversarial** | 0.1 | 0.001 | 0.2 | 80.62% | 7.05% | 39.03% |
| | **0.1** | **0.005** | **0.2** | **80.66%** | **7.90%** | **55.54%** |
| | 0.1 | 0.05 | 0.2 | 80.04% | 7.37% | 78.83% |
| | 0.1 | 0.1 | 0.2 | 80.36% | 8.08% | 83.72% |
| | 0.1 | 0.2 | 0.2 | 80.22% | 8.05% | 87.26% |

- Non-adversarial model: $\lambda_A = 0$, $\lambda_N = 0$, $r_N = 0$
- Adversarial network: $\lambda_A = 0.1$, $\lambda_N = 0$, $r_N = 0$
- Disentangled adversarial network: $\lambda_A = 0.1$, $\lambda_N = 0.005$, $r_N = 0.2$

# 3. Experimental Evaluation and Results: Results and Discussion




- Non-adversarial model:

  $\lambda_A = 0, \lambda_N = 0, r_N = 0$

- Adversarial network:

  $\lambda_A = 0.1, \lambda_N = 0, r_N = 0$

- Disentangled adversarial network:

  $\lambda_A = 0.1, \lambda_N = 0.005, r_N = 0.2$

# Thank you.