2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-6819-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBM55620.2022.9995340

Towards Generalized ML Model in Automated Physiological Arousal Computing: A Transfer Learning-Based Domain Generalization Approach

Ruijie Fang
ECE Department
University of California, Davis
CA, USA
rjfang@ucdavis.edu

Ruoyu Zhang
ECE Department
University of California, Davis
CA, USA

Elahe Hosseini
ECE Department
University of California, Davis
CA, USA

Anna M. Parenteau
Department of Psychology
University of California, Davis
CA, USA

Sally Hang
Department of Psychology
University of California, Davis
CA, USA

Setareh Rafatirad
ECE Department
University of California, Davis
CA, USA

Camelia E. Hostinar
Department of Psychology
University of California, Davis
CA, USA

Mahdi Orooji
ECE Department
University of California, Davis
CA, USA

Houman Homayoun
ECE Department
University of California, Davis
CA, USA

***Abstract*—Physiological signal-based pattern recognition has progressed significantly, such as automated pain assessment and stress detection. Public datasets provide a research platform to conduct machine learning studies. However, models trained from public datasets easily overfit that specific dataset and do not apply to unseen data collected in real-life scenarios. This paper proposes to use the transfer learning-based domain generalization technique to generalize the models to solve this issue. Data from different training domains are generalized, i.e., the dissimilarity is minimized by the proposed approach such that the model trained is generalized. We proved that the generalized model is more adaptive to new unseen data. Experiments have been done on the BioVid heat pain dataset and WESAD stress dataset, and results showed that our proposed methods significantly improve the model performance on new unseen data.**

***Index Terms*—Affective Computing, Domain Generalization, Transfer Learning, Automated Pain Assessment**

I. Introduction

Physiological arousal is an essential indicator of body status [1], [2]. Pain, stress, and emotion are psychological processes that correspond to changes in physiological arousal [3]. Pain, often caused by tissue injuries like heat or electricity, induces nervous system responses and involves changes in physiological arousal, including increases in heart rate and skin conductivity [2]. Similarly, stress activates the sympathetic-adrenal-medullary (SAM) system, whose activity results in changes in peripheral physiology, including heart rate and respiration rate. Historically, emotions have been conceptualized as varying across two dimensions: arousal and valence. Therefore, studying physiological arousal as it relates to perceptions of pain, stress, and emotion is likely to yield a new understanding of the inter-relations among these constructs [1].

It is essential to assess physiological arousal events like pain and stress as severe consequences can happen if they are not detected and well-treated. Pain is the primary reason people visit a doctor [4], which brings undesired feelings, loss of productivity, and financial costs. Overtreatment of pain can lead to drug addiction, while undertreatment of pain can cause increased blood pressure and heart rate. The National Health Interview Survey (NHIS) reported that $296 million loss of productivity in the year 2019 in the U.S. [5]. Besides, stress is a global, common psychological issue that contributes to insomnia, depression, gastrointestinal problems, and multiple stress-related disorders.

As these events are subjective and traditional assessment methods are time-consuming, e.g., the Numerical Rating Scale as a self-report pain assessment method, researchers have conducted various experiments to seek the machine learning-based automated pain/stress/emotion detection methods [2], [6]–[9]. Walter et al. [10] published BioVid dataset in which they used induced thermal pain as the pain stimuli on 90 healthy subjects and recorded facial expression, physiological signals, including skin conductance level (SCL), electrocardiogram (ECG), electromyogram (EMG), and electroencephalography (EEG). Lopez et al. [11] obtained 82.75% accuracy for multi-task classification on the BioVid dataset. Werner et al. published the X-ITE database collected on 134 healthy volunteers while applying thermal and electrical pain and recording video, audio, and physiological signals including EMG, ECG, EEG, and SCL. A range from 83.3% to 94.3% accuracies for different

pain types were reported. Werner et al. [2] summarized the progress in automated pain assessment. In terms of stress, Schmidt et al. [12] published the WESAD dataset that contains physiological signals of ECG, SCL, EMG, Respiration, temperature and acceleration collected from 15 subjects while performing trier social stress test (TSST) experiments and the highest accuracy of 92.83% of stress/non-stress classification was presented. DEAP (A Database for Emotion Analysis using Physiological Signals) recruited 32 participants and used video clips to stimulate participants' arousal, valence, dominance, liking and familiarity [3], [13]. 32-channel EEG, GSR, BVP, RSP, skin temperature, EMG, and EOG were collected.

Although numerous datasets were published and experimental physiological arousal studies results have shown promising results, real-life scenarios were seldom considered [14]. The first issue in real-life scenarios is the newly found disease. In such cases, previously collected datasets do not fit the new disease. Secondly, in real-life scenarios, many factors can change, for example, the use of sensors, devices, the different groups of subjects and the actual physiological arousal events versus the experimental induced physiological arousal. These factors can contribute to the change in data distribution [15]. Suppose models trained on publicly available datasets are directly applied to newly collected data; the results may degrade because the pre-trained model might be overfitting on its training dataset and freshly collected data have a different distribution [16]. The most obvious and efficient way to overcome this issue is to collect labeled data with the same settings to ensure the same data distribution, but it is time-consuming and human resource-consuming. Last but not least, supervised learning faces the issue of a "cold start," which means that at the beginning, with a small number of the training set, the classifier performance tends to be poor. As a new, rapidly progressing field in machine learning, transfer learning offers us another option to deal with these issues. Transfer learning aims to use the knowledge from the source domain on the target domain, where a domain is the feature space and its data distribution [17]. Theerthagiri [18] proposed a new stress emotion recognition method based on CNN and transfer learning, but the proposed SER method is supervised and does not apply to new unseen data. In another work by Kachele et al. [19], an ensemble classifier-based regression system was proposed to measure the confidence of samples, and then, only samples with high confidence would be added to the training set. However, their methods only used the part of "similar" data samples and abandoned the rest of the entire dataset, which missed usable information hidden in the abandoned samples. Li et al. [20] extracted differential entropy features and employed Adversarial domain adaptation + association reinforcement to classify valence and arousal on DEAP and SEED datasets. Zhang et al. [21] deployed TrAdaboost to recognize valence and arousal on the DEAP dataset and gained cross-subject accuracy of 66.7% and 66.1%. However, TrAdaboost is an inductive transfer learning algorithm, i.e., it requires a small amount of labeled data which makes it limited to cases with labeled new domain data.

We propose using transfer learning, specifically domain generalization techniques, to train a generalized model from two datasets, BioVid and WESAD [10], [12]. The generalized model takes input data from both datasets and minimizes the dissimilarity between these two datasets by deploying transfer learning algorithms. Then, to simulate the real-life scenario, we assume the newly collected data have the same distribution as the combination of data from the BioVid heat pain dataset and WESAD stress dataset [22]. Under this assumption, we randomly pick one-third of subjects from BioVid and one-third of subjects from WESAD and combine them to form the testing set and ensure they are unseen in the training set. Note that a primary prerequisite to combining these two datasets is that they share the same biosignal data, which will be further discussed in Section II. We conducted four experiments, including (1) directly applying experiment where we directly applied the model training from two datasets on the test subjects, (2) normalization per subject experiment where instead of normalizing the entire dataset, we normalize the data within each subject, (3) feature engineering experiment where we performed a series of feature engineering work to increase the performance of the classifier and (4) DICA experiment where we deployed domain invariant component analysis (DICA) which is a domain generalization technique. Finally, a support vector machine and random forest were deployed to train the models. The results showed that the classification performance for these four series experiments increased gradually from 67% to 83%, 88% and at the end, 91.8% when DICA was applied. The framework of the proposed approaches is illustrated in Fig. 1.

The main contributions of this paper can be summarized as follows:

- We first consider physiological arousal as a whole and propose to train generalized models to solve the issue of overfitting to the specific dataset and "cold start".
- We propose a series of techniques to optimize the generalized model training including normalization per subject, feature engineering and DICA.
- We conducted experiments on the BioVid pain and WESAD stress datasets to verify the idea. Results showed that our proposed method helped the classifier grow from 67% in accuracy to 91.8% accuracy at the end.

The remainder of this paper is organized as follows. In Section II, we give a brief introduction of the dataset used and a detailed description of the proposed approaches. Then, in Section III, we present the experiments we conducted and the evaluation results of the experiments. Following that, Section IV contains the merits, potential and weaknesses of this study and the future directions to go and explain the reasons of choosing datasets, algorithms in this study. Lastly, Section V gives the findings of this study.

II. Method

A. Dataset

The experiments are conducted on the BioVid heat pain and WESAD stress datasets. The BioVid dataset is a public

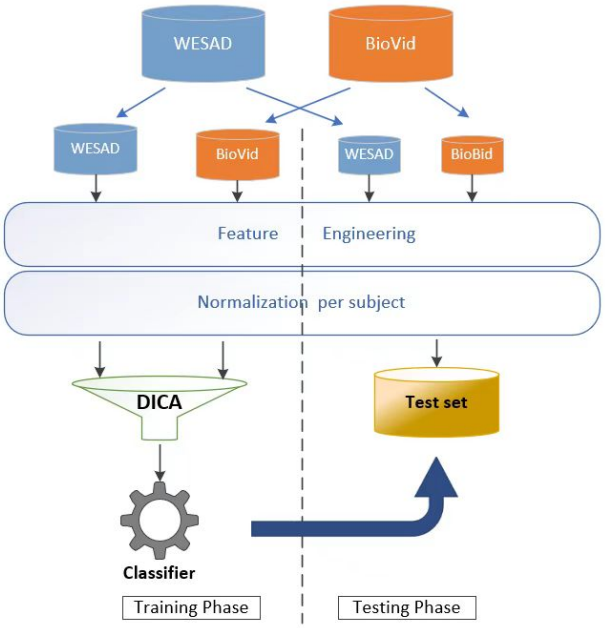


Fig. 1. The framework of the proposed series approaches.

database that contains both biosignals and video data, facial video, skin conductivity level (SCL), electrocardiogram (ECG) and sEMG (from trapezius muscle, corrugator and zygomaticus) data with induced heat pain of four different pain levels on 87 healthy volunteers. WESAD dataset focuses on stress while having similar data collection methods. It also has ECG, SCL and EMG from trapezius. Further details can be found in [10] and [12].

As illustrated in Section I, one prerequisite of training a generalized model is that all training datasets should have the same signal configurations. This is because the technique of domain generalization assumes that all training domains share the same feature space. Thus, we pick ECG, SCL, and sEMG of the trapezius, which are the same biosignals that these two datasets share.

B. Fundamental framework

As the basement of this study and verifying if the issue of models overfitting specific datasets exists, we propose this fundamental framework. First, datasets are split into a training set and a testing set in the manner of 7:3 and merged into a whole training set and a testing set as shown in Fig. 1. It is done by subjects instead of samples, i.e., WESAD has 15 subjects, and ten subjects will form the training set while the remaining five subjects form the testing set. This way, the potential overfitting issue caused by data from the same subject is eliminated. Then, a list of commonly used features is extracted as shown in Table I. Features indexed 1 to 16 are used for SCL and sEMG, while features indexed 17 to 26 are used for ECG. Thus a total of 42 features are used. In the end, the support vector machine (SVM) with "RBF" kernel and random forest (RF) were trained on the training set and tested on the testing set [33], [34]. The prediction output is high/low physiological arousal from both datasets instead of single pain or stress as we consider them as a whole.

C. Normalization per subject

Because the training set has two data sources: BioVid and WESAD, and these two datasets are collected with different devices on different subjects group, we hypothesize there exists a dissimilarity in the data distribution of the two datasets. In addition, within each dataset, data have different distributions among different subjects because individuals have different baseline vital signs and different responses to pain or stress. To reduce such dissimilarity, we propose to normalize data per subject. Normalization is transforming data into the range [0, 1] (or any other range) or simply transforming data onto the unit sphere. It is necessary for some scenarios, especially when Euclidean distance is used, e.g., if the classifier is SVM and is not needed sometimes. Traditionally, it is done on the entire dataset, which neglects the dissimilarity among subjects. The normalization equation is illustrated in equation 1.

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{1}$$

We propose to normalize data per subject as, in this way, the dissimilarity can be initially reduced. Fig. 2 and Fig. 3 shows how two normalization works and the differences between traditional normalization and normalization per subject. We also conducted experiments to verify the effectiveness of this approach which will be further discussed in Section III.

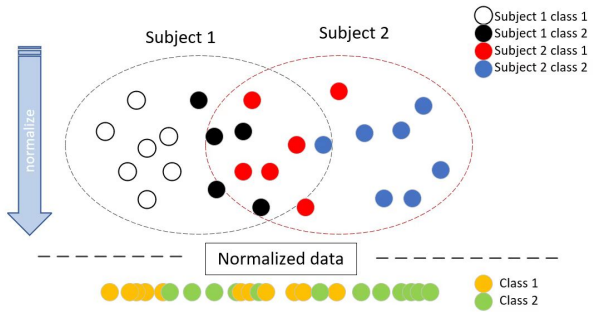


Fig. 2. The traditional normalization approach.

D. Feature engineering

Feature extraction is of great importance in machine learning as it transforms implicit, non-discriminative data into informative and discriminative data. How discriminative the data is mainly depending on the feature extracted. Table I summarized all features we have extracted in a total of 42 features. Features indexed 1 to 16 are used for SCL and sEMG, while features indexed 17 to 26 are used for ECG. For detailed information on these features, please refer to [23], [31]. The feature engineering aims to remove redundant or non-informative features. We used three feature extraction techniques: variance threshold, SelectKbest, and tree-based

TABLE I
EXTRACTED FEATURE LIST

Index	Feature	Description	References
1	MAV	Mean Absolute Value	[6], [23]
2	P2P	Peak to Peak Amplitude	[6], [23]
3	Peak	Peak Amplitude	[6], [23]
4	SD	Standard Deviation	[6], [23]
5	VAR	Variance	[23], [24]
6	Range	Range	[23], [25]
7	IQR	Interquartile Range	[23]
8	MeanFreq	Mean Frequency	[23]
9	MedianFreq	Median Frequency	[23]
10	ModeFreq	Mode Frequency	[23]
11	ZeroCrossings	Zero Crossings	[23]
12	ApEn	Approximate Entropy	[23], [26]
13	FuzzyEn	Fuzzy Entropy	[23], [27]
14	SampEn	Sample Entropy	[23], [28]
15	ShannonEn	Shannon Entropy	[23], [29]
16	SpectralEn	Spectral Entropy	[23], [30]
17	MeanNN	Mean NN interval	[23], [31]
18	SDNN	Standard deviation of NN intervals	[23], [31]
19	RMSSD	Root mean square of successive RR interval differences	[23], [31]
20	SDSD	Standard Deviation of Standard Deviation Vector	[31]
21	CVNN	Coefficient of Variation of NN intervals	[31]
22	MedianNN	Median NN interval	[31]
23	IQRNN	Interquartile range of NN interval	[31]
24	pNN50	Percentage of successive RR intervals differ by more than 50 ms	[31], [32]
25	pNN20	Percentage of successive RR intervals differ by more than 20 ms	[31]
26	TINN	Baseline width of the RR interval histogram	[31]

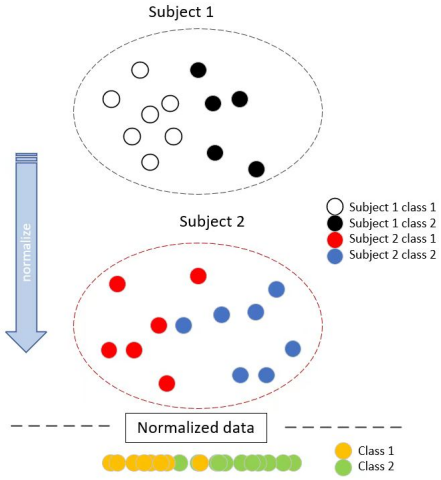


Fig. 3. The normalization per subject approach.

selecter. The variance threshold is a primary method of feature selection that removes all those features whose variance does not meet some threshold. SelectKBest is a single feature selection approach that selects the best features through a univariate statistical test. It can be used as a preprocessing step for the evaluator. SelectKBest, precisely, removes all but the top K-ranked features. Tree-based selecter trains several tree-based classifiers and applies them to the dataset. The tree models will generate feature importance scores based on each feature's entropy and information gain and classification results.

E. DICA

The research objective of Domain Generalization is figuring out how to train a model with source domain data, whose purpose is to make the model generalize to the target domain with different data distributions, as shown in Fig. 4. Although this ability does not seem to require deliberate learning for humans, domain generalization is still a very challenging problem for current machine learning models and algorithms. According to [35], domain generalization is defined as:

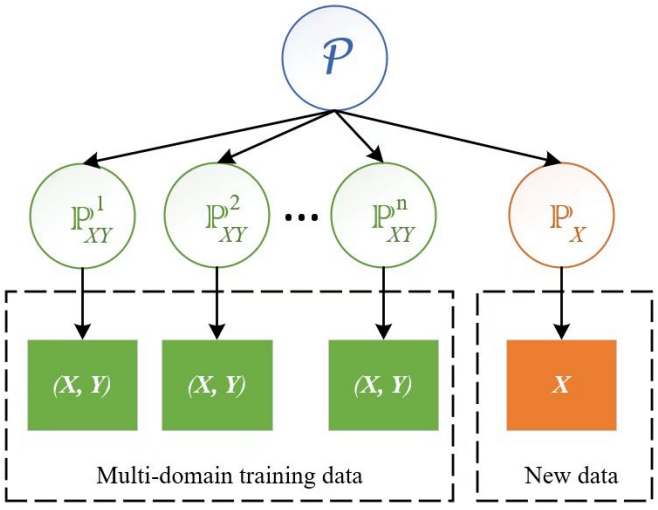


Fig. 4. Domain Generalization Schematic

The testing data come from N domains with different but similar data distribution $\mathcal{S} = \mathcal{S}_{i\,i=1}^{N}$. Within each domain, data

obey its distribution $(\boldsymbol{x}_j, y_j)$ $P_i(\boldsymbol{x}, y)$*. Domain generalization learns a model* $f : \boldsymbol{x} \to \mathbb{R}$ *such that* f *minimizes the prediction error on the testing set* $\mathcal{S}_t$ *formed from* N *domains.*

We propose to use a data-based domain generalization method DICA, which is short for domain invariant component analysis [36]. Data-based domain generalization methods aim to minimize the data distribution among several source domains, learn models independent of domains and better adapt to new data. DICA, specifically, derived from the idea of transfer component analysis (TCA) [37], tries to find a transformation such that in this transformer space, the dissimilarity of data distribution from all data is the minimum. The dissimilarity is defined as *Distributional Variance*:

$$\mathbb{V}_{\mathcal{H}}(\mathcal{P}) := \frac{1}{N}\operatorname{tr}(\Sigma) = \frac{1}{N}\operatorname{tr}(G) - \frac{1}{N^2}\sum_{i,j=1}^{N} G_{ij}, \quad (2)$$

where N represents the number of domains. $\boldsymbol{\Sigma}$ is the covariance operator of $\mathcal{P}$ which is defined as:

$$\Sigma := G - 1_N G - G 1_N + 1_N G 1_N, \quad (3)$$

where $\mathbf{1}_N$ represents all-1 matrix with size of N and matrix $\mathbf{G}$ is the Gram matrix of sample inner product.

In order to calculate VH, DICA makes an empirical representation:

$$\widehat{\mathbb{V}}_{\mathcal{H}} = \frac{1}{N}\operatorname{tr}(\widehat{\Sigma}) = \operatorname{tr}(KQ), \quad (4)$$

where $\mathbf{K}$ and $\mathbf{Q}$ are the block kernel and coefficient matrices.

Thus, we can seek a transform matrix $\mathbf{B}$ after which the distribution variance is minimized. The empirical distributional variance between sample distributions is:

$$\min \operatorname{tr}\left(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{Q}\boldsymbol{K}\boldsymbol{B}\right). \quad (5)$$

By combining these two targets, the ultimate optimization problem becomes:

$$\max_{\boldsymbol{B}} \frac{\frac{1}{n}\operatorname{tr}\left(B^{\mathrm{T}} L \left(L + n\varepsilon I_n\right)^{-1} K^2 B\right)}{\operatorname{tr}\left(B^{\mathrm{T}} K Q K B + B K B\right)}. \quad (6)$$

Matrix $\mathbf{B}$ is the goal of DICA. The data distribution dissimilarity is minimized if the minimized $\mathbf{B}$ is found and applied.

III. Experiments and Results

In this section, we present the four experiments, including (1) directly applying experiment where we directly applied the model training from two datasets on the test subjects, (2) normalization per subject experiment where we verified the performance of the normalizing per subject approach as illustrated in Section II, (3) feature engineering experiment where we performed a series of feature engineering work (variance threshold, SelectKBest, and tree-based selecter) to remove undesired features and (4) DICA experiment where we deployed DICA and evaluate its effectiveness.

TABLE II
Classification results of directly applying experiment

Dataset	Model	Accuracy	F-1 Score	AUC
BioVid	RF	77.2%	76.1%	0.78
WESAD	RF	97.2%	97.5%	0.99
BioVid	SVM	75.9%	77.0%	0.79
WESAD	SVM	96.7%	96.3%	0.98
Multi-domain	RF	62.8%	64.1%	0.67
Multi-domain	SVM	62.3%	64.7%	0.65

A. Directly Applying Experiment

In this experiment, we aim to verify the hypothesis that the pre-trained model performs worse on the newly collected dataset because of the overfitting of a specific dataset. First, we trained SVM and RF models on each dataset and cross-validate them separately. The performance of these models can be used as a baseline reference. Table II displayed the evaluation results of such experiment and Fig. 5 (a) and (b) present the receiver operating characteristic curves of RF models.

Next, we tested the performance of pre-trained models on the new dataset. Around 70% of subjects from either BioVid and WESAD form the training dataset the combination of the resting 30% of both BioVid and WESAD form the testing set. The RF classification yielded 60.3% accuracy, 0.62 AUC and 68.8% accuracy, and 0.71 AUC for the BioVid-trained and WESAD-trained models, respectively.

Then, we conducted a multi-domain training in which around 70% of subjects from BioVid and WESAD form the multi-domain training domain, and the rest 30% subjects form the testing domain. We use the same feature extraction technique and machine learning algorithms to train RF and SVM models, and the results are presented in table II and the ROC of the RF model is shown in Fig. 5 (c). The results yield that all multi-domain training performance evaluation metrics are poorer than traditional machine learning tasks.

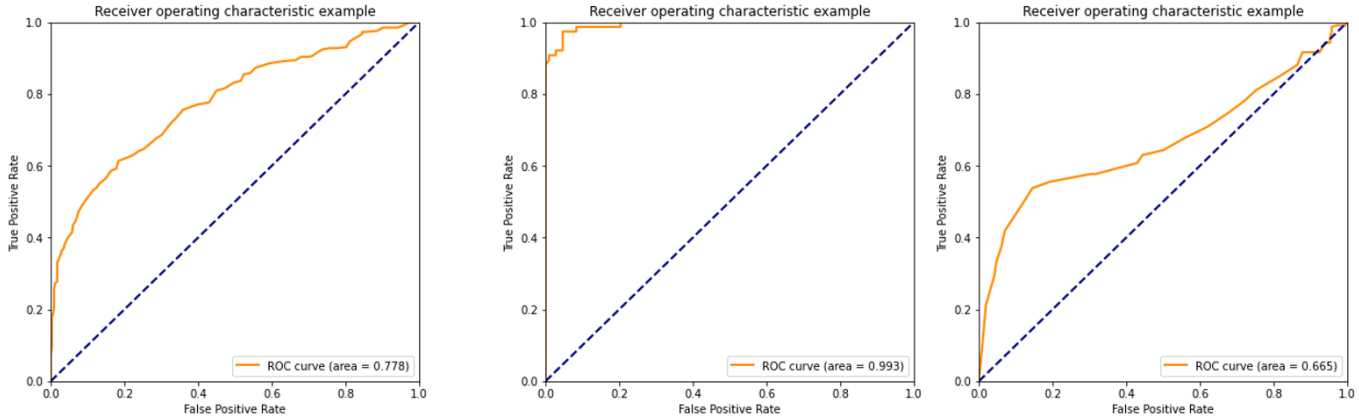
B. Normalization per Subject Experiment

In this experiment, we normalized the data per subject and conducted the same multi-domain machine learning pipeline as in previous experiments to evaluate the influence of normalization per subject. The evaluation metrics results showed an accuracy of 83.2%, F-1 score of 82.8% with AUC of ROC as 0.88 from the RF model and 80.4% accuracy, 81.7% f-1 and AUC of 0.86 from SVM model and ROC of RF is shown in Fig. 6 (a).

C. Feature Engineering Experiment

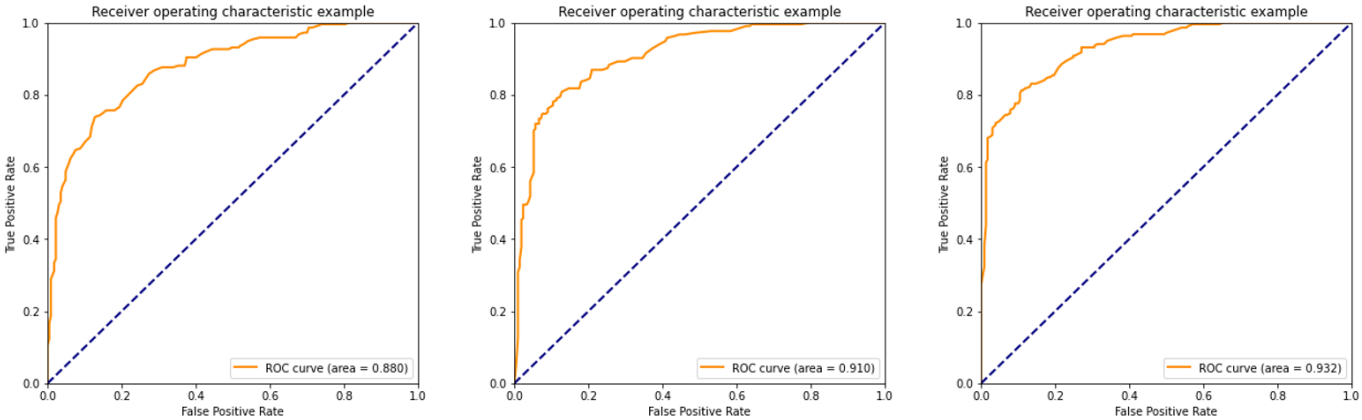
In this section, we deployed the feature engineering techniques discussed based on the normalization per subject settings. For the variance threshold method, the threshold is set to be 0.1, which filters the feature number from 42 to 27. Then, in the SelectKBest selection, we select the top 15 features, and finally, in the tree-based selecter, ten features were finalized. The list of finalized features is shown in table III.

After features were finalized, the same multi-domain machine learning pipeline was deployed again to verify the effectiveness of feature engineering. Results showed an accuracy



(a) BioVid (b) WESAD (c) Multi-domain

Fig. 5. ROC of RF models in directly applying experiments.



(a) Normalization per subject (b) Feature engineering (c) DICA

Fig. 6. ROC of RF models.

TABLE III
FINALIZED FEATURE LIST FROM FEATURE ENGINEERING EXPERIMENT

EDA_MAV	EDA_P2P	EDA_RMS	EDA_VAR
EDA_ModeFreq	EMG_MAV	EMG_SD	HRV_MeanNN
HRV_pNN50	HRV_pNN20		

of 87.8%, F-1 score of 88.7% with AUC of ROC as 0.91 from the RF model and 88.9% accuracy, 87.0% F-1 and AUC of 0.89 from the SVM model. The ROC of RF model is displayed in Fig. 6 (b).

D. DICA Experiment

In this experiment, we deployed all techniques discussed above and applied DICA to the multi-domain training set. The entire pipeline is presented in Fig. 1. The evaluation metrics are shown in table IV and the ROC is presented in Fig. 6 (c).

threeparttable

IV. DISCUSSION

In this section, we discuss the merits, weaknesses and potential future directions of this study.

TABLE IV
CLASSIFICATION RESULTS OF DICA EXPERIMENT

Experiment	Model	Accuracy	F-1 Score	AUC
NPS	RF	82.2%	82.8%	0.88
NPS	SVM	80.4%	81.7%	0.86
FEE	RF	87.8%	88.7%	0.91
FEE	SVM	88.9%	87.0%	0.89
DICA	RF	91.8%	92.1%	0.932
DICA	SVM	91.2%	91.3%	0.92

NPS refers to normalization per subject experiment
FEE refers to feature engineering experiment

Three experiments were conducted in the directly applying experiments, including BioVid and WESAD training and testing on their own, respectively, and a multi-domain experiment. Table II and Fig. 5 presented the results revealed the performance deduction when multi-domain was applied. Such deduction verified our hypothesis that the pre-trained model is overfitting to a specific dataset and is unready to be used in real-life scenarios. A similar phenomenon can also be found in [16]. In automated pain/stress/emotion detection, numerous

datasets have been published, and researchers have achieved promising results with the datasets collected in experimental settings. However, when real-life scenarios are considered, the fancy buildings seem to be a landslide. Such performance deduction reminds us of the urgent need for clinical data collection and advanced data adaptation techniques such as transfer learning.

We propose to use normalization per subject as we hypothesize that different individuals' data may have different distributions, and if they are normalized directly, such dissimilarity persists and results in poor discriminative ability as shown in Fig. 2. When normalization per subject is applied, the difference in distribution is undermined, as shown in Fig. 3. Experiments have been conducted to verify this assumption, and results show that with normalization per subject, the classification performance significantly increased in an average accuracy increment of 19.3%. This shocking increment supports the idea that dissimilarity in data distribution exists in the different individuals' physiological signals, and the proposed method can effectively weaken it. Thus, we have evidence to support this approach when dealing with multiple subjects' data. However, if such an approach is used in clinical settings for new subjects, there is still the challenge that it takes some time to collect enough data samples to get to know a new subject's data distribution before it is ready to normalize. This can also be regarded as a 'cold start' problem; thus, the amount of sufficient data samples is a worthwhile direction to pursue in the future.

The feature engineering experiment results showed the effectiveness of the proposed approach. Due to the restriction of multiple datasets involved, some features are not able to be used, such as frequency domain features for ECG signal as the BioVid dataset applied experimental pain for 5.5 seconds and is insufficient for frequency domain features calculation as they require a minimum of 60-second data. However, It is inhuman to induce pain for a long duration. Thus, there is a demand for more clinical data research.

In terms of transfer learning, We chose DICA as the domain generalization solution of the proposed problem because of the following reasons: among various domain generalization methods, DICA is classic that it is based on a general and practical theory and numerous latest domain generalization algorithms are derived from it but not generalized as it is and may not perform better than DICA [38]. In addition, DICA assumes $P(Y|X)$ is stable while distribution shifts only happen to $P(X)$. We believe this assumption best represents the case for affective computing since the universal responses of pain/stress/emotion [7]. However, due to the lack of studies using domain generalization methods in affective computing, other domain generalization algorithms are also worth trying in future studies; for example, other than DICA, there has been a way of minimizing conditional possibility distribution which was the conditional-invariant domain generalization (CIDG) approach [39] and also Fisher-like discriminative analysis approach scatter component analysis (SCA) [40]. These methods used different approaches and were worthy of testing on physiological datasets.

When DICA was applied, the classification accuracy increased from 87.8%, 88.9% to 91.8%, 91.2% for RF and SVM, respectively. Such promising results showed that there is a dissimilarity between WESAD and BioVid datasets, possibly due to different experiment tasks, sensors used, and different groups of people. By applying DICA, an average of 3.15% accuracy increment was achieved, showing that such data-based domain generalization methods can weaken the dissimilarity and find the joint distribution among different domains. This is consistent with previous work [16] that transfer component analysis (TCA) can be used to minimize dissimilarity between the source domain and target domain.

Furthermore, we use two machine learning models in this study, random forest and support vector machine, to verify the performance of the proposed framework since they are classic and most commonly used in pain detection and emotion detection [3]. Since the proposed method is a framework, it can adopt any fundamental machine learning algorithm to replace SVM and RF, and more machine learning models are encouraged to plug in, for example, ensemble learning (e.g., Adaboost) and deep learning.

In terms of the data, we used data from the BioVid heat pain dataset and WESAD stress dataset and treated pain and stress together because of the mechanism that pain and stress share as the model of physiological arousal. In addition, because DICA assumes all training domains share the same feature space, both BioVid and WESAD have the same modalities of ECG, SCL, and EMG data. To further investigate the proposed hypothesis and solution, we intend to include more datasets in the future. However, when more datasets are involved, this prerequisite will be broken, and an alternative approach needs to be deployed to overcome the differences among datasets.

V. CONCLUSION

This paper proposes a transfer learning-based domain generalization framework to train a generalized model that fits newly collected data from real-life scenarios. The proposed framework contains three major components: (1) normalization per subject, (2) feature engineering, and (3) DICA. DICA is the domain generalization algorithm that defines a distributional variance that describes the cross-domain variance and aims to minimize this variance to eliminate the dissimilarity across multi-domains. Four experiments have been conducted to verify the success of each component in the proposed system. Step by step, the classification accuracy increased from an initial 62.8% to 91.8% at the end, which indicates. Although these experiments yielded promising results, there were limitations. For example, there was a lack of emotional data. Future studies should include more datasets, a different combination of sensors, and other transfer learning algorithms.

REFERENCES

[1] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. A review of emotion recognition using physiological signals. *Sensors*, 18(7):2074, 2018.

[2] Philipp Werner, Daniel Lopez-Martinez, Steffen Walter, Ayoub Al-Hamadi, Sascha Gruss, and Rosalind Picard. Automatic recognition methods supporting pain assessment: A survey. *IEEE Transactions on Affective Computing*, 2019.

[3] Wei Li, Zhen Zhang, and Aiguo Song. Physiological-signal-based emotion recognition: An odyssey from methodology to philosophy. *Measurement*, 172:108747, 2021.

[4] Pekka Mäntyselkä, Esko Kumpusalo, Riitta Ahonen, Anne Kumpusalo, Jussi Kauhanen, Heimo Viinamäki, Pirjo Halonen, and Jorma Takala. Pain as a reason to visit the doctor: a study in finnish primary health care. *Pain*, 89(2-3):175–180, 2001.

[5] R Jason Yong, Peter M Mullins, and Neil Bhattacharyya. Prevalence of chronic pain among adults in the united states. *Pain*, 163(2):e328–e332, 2022.

[6] Sascha Gruss, Roi Treister, Philipp Werner, Harald C Traue, Stephen Crawcour, Adriano Andrade, and Steffen Walter. Pain intensity recognition rates via biopotential feature patterns with support vector machines. *PloS one*, 10(10):e0140330, 2015.

[7] Philip Schmidt, Attila Reiss, Robert Dürichen, and Kristof Van Laerhoven. Wearable-based affect recognition—a review. *Sensors*, 19(19):4079, 2019.

[8] Zhichao Zhang, Ruoyu Zhang, Chi-Wei Chang, Yaojun Guo, Yung-Wei Chi, and Tingrui Pan. iwrap: A theranostic wearable device with real-time vital monitoring and auto-adjustable compression level for venous thromboembolism. *IEEE Transactions on Biomedical Engineering*, 68(9):2776–2786, 2021.

[9] Gozde Goncu-Berk, Ruoyu Zhang, and Cigdem Yilmaz. *CalmWear: A Smart Tactile Sensory Stimulation Clothing*, page 184–188. Association for Computing Machinery, New York, NY, USA, 2021.

[10] Steffen Walter, Sascha Gruss, Hagen Ehleiter, Junwen Tan, Harald C Traue, Philipp Werner, Ayoub Al-Hamadi, Stephen Crawcour, Adriano O Andrade, and Gustavo Moreira da Silva. The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In *2013 IEEE international conference on cybernetics (CYBCO)*, pages 128–131. IEEE, 2013.

[11] Daniel Lopez-Martinez and Rosalind Picard. Multi-task neural networks for personalized pain recognition from physiological signals. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 181–184. IEEE, 2017.

[12] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018.

[13] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

[14] Ruijie Fang, Ruoyu Zhang, Sayed M Hosseini, Mahya Faghih, Soheil Rafatirad, Setareh Rafatirad, and Houman Homayoun. Pain level modeling of intensive care unit patients with machine learning methods: An effective congeneric clustering-based approach. In *2022 4th International Conference on Intelligent Medicine and Image Processing*, pages 89–95, 2022.

[15] Seyed Ali Rokni, Marjan Nourollahi, Parastoo Alinia, Iman Mirzadeh, Mahdi Pedram, and Hassan Ghasemzadeh. Transnet: minimally supervised deep transfer learning for dynamic adaptation of wearable systems. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, 26(1):1–31, 2020.

[16] Ruijie Fang, Ruoyu Zhang, Elahe Hosseini, Sayed Hosseini, Mahya Faghih, Mahdi Orooji, Soheil Rafatirad, Setared Rafatirad, and Houman Homayoun. Atlas: An adaptive transfer learning based pain assessment system: A real life unsupervised pain assessment solution. *44th International Engineering in Medicine and Biology Conference*, 2022.

[17] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[18] Prasannavenkatesan Theerthagiri. Stress emotion recognition with discrepancy reduction using transfer learning. *Multimedia Tools and Applications*, pages 1–15, 2022.

[19] Markus Kächele, Mohammadreza Amirian, Patrick Thiam, Philipp Werner, Steffen Walter, Günther Palm, and Friedhelm Schwenker. Adaptive confidence learning for the personalization of pain intensity estimation systems. *Evolving Systems*, 8(1):71–83, 2017.

[20] Jinpeng Li, Shuang Qiu, Changde Du, Yixin Wang, and Huiguang He. Domain adaptation for eeg emotion recognition based on latent representation similarity. *IEEE Transactions on Cognitive and Developmental Systems*, 12(2):344–353, 2019.

[21] Xiaowei Zhang, Wenbin Liang, Tingzhen Ding, Jing Pan, Jian Shen, Xiao Huang, and Jin Gao. Individual similarity guided transfer modeling for eeg-based emotion recognition. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1156–1161. IEEE, 2019.

[22] Kayode Peter Ayodele, Wisdom O Ikezogwo, Morenikeji A Komolafe, and Philip Ogunbona. Supervised domain generalization for integration of disparate scalp eeg datasets for automatic epileptic seizure detection. *Computers in Biology and Medicine*, 120:103757, 2020.

[23] Evan Campbell, Angkoon Phinyomark, and Erik Scheme. Feature extraction and selection for pain recognition using peripheral physiological signals. *Frontiers in neuroscience*, 13:437, 2019.

[24] Angkoon Phinyomark, Pornchai Phukpattaranont, and Chusak Limsakul. Feature reduction and selection for emg signal classification. *Expert systems with applications*, 39(8):7420–7431, 2012.

[25] Steffen Walter, Sascha Gruss, Kerstin Limbrecht-Ecklundt, Harald C Traue, Philipp Werner, Ayoub Al-Hamadi, Nicolai Diniz, Gustavo Moreira da Silva, and Adriano O Andrade. Automatic pain quantification using autonomic parameters. *Psychology & Neuroscience*, 7(3):363–380, 2014.

[26] Steven M Pincus. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6):2297–2301, 1991.

[27] Bart Kosko. Fuzzy entropy and conditioning. *Information sciences*, 40(2):165–174, 1986.

[28] Joshua S Richman and J Randall Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 2000.

[29] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[30] Aihua Zhang, Bin Yang, and Ling Huang. Feature extraction of eeg signals using power spectral entropy. In *2008 international conference on BioMedical engineering and informatics*, volume 2, pages 435–439. IEEE, 2008.

[31] Fred Shaffer and Jay P Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in public health*, page 258, 2017.

[32] Mingzhe Jiang, Riitta Mieronkoski, Amir M Rahmani, Nora Hagelberg, Sanna Salanterä, and Pasi Liljeberg. Ultra-short-term analysis of heart rate variability for real-time acute pain monitoring with wearable electronics. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1025–1032. IEEE, 2017.

[33] Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[34] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[35] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24, 2011.

[36] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013.

[37] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE transactions on neural networks*, 22(2):199–210, 2010.

[38] Shoubo Hu, Kun Zhang, Zhitang Chen, and Laiwan Chan. Domain generalization via multidomain discriminant analysis. In *Uncertainty in Artificial Intelligence*, pages 292–302. PMLR, 2020.

[39] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[40] Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1414–1430, 2016.