2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871536

ATLAS: An Adaptive Transfer Learning Based Pain Assessment System: A Real Life Unsupervised Pain Assessment Solution

Ruijie Fang[1], Ruoyu Zhang[1], Elahe Hosseini[1], Sayed Mohammad Hosseini[2], Mahya Faghih[2], Mahdi Orooji[1], Soheil Rafatirad[3], Setareh Rafatirad[4] and Houman Homayoun[1]

Abstract— **Undertreatment or overtreatment of pain will cause severe consequences physiologically and psychologically. Thus, researchers have made great efforts to develop automatic pain assessment approaches based on physiological signals using machine learning techniques. However, state-of-art research mainly focuses on verifying the hypothesis that physiological signals can be used to assess pain. The critical assumption of these studies is that training data and testing data have the same distribution. However, this assumption may not hold in real-life scenarios, for instance, the adoption of machine learning model by a new patient. Such real-life scenarios in which user's data is unlabeled is largely neglected in literature. This study compensates for the rift by proposing an adaptive transfer learning based pain assessment system (ATLAS), a novel adaptive learning system based on the transfer learning algorithm Transfer Components Analysis (TCA) to minimize the distance between training data and unlabeled testing data. Experiments were conducted on BioVid database, and the results showed our approach outperforms three existing traditional machine learning-based approaches and achieves an accuracy just 2.0% below the accuracy with labeled data.**

I. INTRODUCTION

National Health Interview Survey (NHIS) in 2019 reported a total loss of productivity of $296 million due to pain [1]. In addition, one out of five adults is suffering from chronic pain. Pain comes with daily life activity interruption, financial loss and mental suffering. Furthermore, undertreatment of pain causes psychological torture and physiological consequences, e.g., increased blood pressure and heart rate. On the other hand, overtreatment of pain may result in nausea, vomiting, or constipation immediately and drug addiction in the long term [2]. To achieve better pain management, pain assessment, as the first step of pain management, needs to be accurate, regular, universal and objective. The traditional self-report methods such as Visual Analogue Scale (VAS), Numerical Rating Scale (NRS) and Verbal Rating Scale (VRS) are falling to disuse as publications are proving that physiological signals, video and audio can help to track an individual's pain level [5]. However, most of the research only focuses on the verification of the hypothesis that the measurements of different modalities can be used to evaluate pain levels. Only a few studies considered testing a pre-trained model on unlabeled data, which is the scenario of using the automatic pain assessment approach in real life. For each new patient, it is resource and time-consuming to do a series of pain validation experiments to get a set of labeled data. In addition, the patient may not be able to assist in such experiments due to their disability to communicate, or the disease pain itself is too hurtful to tell whether pain exists or not. Thus, in real-life cases, the test set is unlabeled but researchers seldomly lmake such important assumption. A typical scenario can be considered as following: A hospital purchased some wearable devices, recruited a certain number of participants and applied induced pain experiments to train a machine learning model based on the collected data. Then, a new patient comes in and the hospital attempts to use the trained model to detect pain levels on this particular patient. However, this patient may not share the same pain response features and the data distribution will be different from the one that the trained model assumed. Thus, the detection performance will be worse due to the difference between the training phase (i.e., the recruited participants) and the test phase (i.e., the new patient). Therefore, it is crucial to develop an algorithm that can adapt to new patients and perform accurately and robustly on new patients' data.

In recent years many researchers developed various algorithms to objectively assess the level of pain [3], [4]. Among those works there are several that are closely related to this work. Kächele et al. [6] trained individual models for each subject using other "similar" subjects' data. The similarity between subjects is measured as 1) the similarity of metadata (i.e., age, gender and questionnaires), 2) the Euclidean distance between data samples, and 3) machine learning (e.g., classification confidence using Random Forest). In another work by Kächele et al. [7], an ensemble classifier-based regression system was proposed to measure the confidence of samples, and then, only sample 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. Chen et al. [8] implemented an inductive transfer learning algorithm, known as "TrAdaboost", which increased the instance weight that is beneficial to target classification tasks. It allowed users to use a small amount of newly labeled data combined with weighted old data to build a high-quality classification model for new data. Nevertheless, TrAdaboost is an inductive transfer learning algorithm, i.e., it requires a small amount

[1]Ruijie Fang, Elahe Hosseini, Ruoyu Zhang, Mahdi Orooji, Houman Homayoun are with the Department of Electrical and Computer Engineering, University of California, Davis, USA rjfang@ucdavis.edu, hhomayoun@ucdavis.edu

[2]Sayed Mohammad Hosseini, Mahya Faghih are with Johns Hopkins Hospital, USA

[3]Soheil Rafatirad is with the Department of Computer Science, University of California, Los Angeles, USA

[4]Setareh Rafatirad is with the Department of Computer Science, University of California, Davis, USA

of data and such a small amount of data may be tough to get in clinical settings.

Transfer learning is a newer learning paradigm in the machine learning field. A key assumption of the traditional machine learning approach is that the test data are drawn from the same distribution as the training data. However, this assumption does not hold in many real-world scenarios [9] including clinical pain management scenarios. Thus, transfer learning is developing rapidly to deal with the rift between testing data and training data. According to the survey paper by Pan et al. [10], Transfer Learning is defined as:

Transfer learning strives to attain better learning performance of prediction function $f_T()$ in the target domain D_T with learning task T_T, using the knowledge in the source domain D_S with learning task T_S, where $D_S \neq D_T$, or $T_S \neq T_T$.

The typical scenario in terms of pain assessment study that we illustrated above can be reused here. Data collected from the recruited experiment participants can be considered source domain D_S, where data is well labeled and data collected from the new patient can be considered as target domain D_T, where data is not labeled. Due to differences in patients' metadata [6] or disease diagnosis, their pain responses are different, which reflects in different probability distributions in the source domain and target domain. The tasks in both source and target domains are the same: to classify pain levels into certain pain levels based on the measured physiological signals.

Transfer learning, based on the availability of labeled data in target domain, can be categorized into three classes: *Inductive Transfer Learning*, where labeled data is available in the target domain, *Transductive Transfer Learning*, where labeled data is available only in the source domain and *Unsupervised Transfer Learning*, where no labeled data in both source and target domain. In automatic pain assessment study, the task mainly lies in transductive transfer learning. Huang et al. [11] proposed to use kernel-mean matching (KMM) algorithm by matching the means in a reproducing-kernel Hilbert space (RKHS) between the source domain and target domain. Dai et al. [12] revised model for target domain based on the traditional Naive Bayes classifiers using EM algorithm. Pan et al. [13] proposed Transfer Component Analysis (TCA) method, which minimizes the distance between the source domain and target domain by using the kernel method to reduce the maximum mean discrepancy (MMD). TCA considers minimizing the distance and preserving data variance, which makes TCA robust and accurate.

This paper proposes ATLAS, an adaptive patient-centered automatic pain assessment system based on the transfer component analysis algorithm. The proposed approach makes use of pre-trained model based on labeled experiment data and can adapt the model based on the new user's unlabeled physiological signal data to improve the classification performance. The main contributions of this paper are as follows:

- We propose an automatic adaptive pain assessment system to improve the performance for unlabeled user.
- We evaluate the classification performance of our proposed system by comparing with three state-of-art methods using traditional machine learning algorithms.

The remainder of this paper is structured as follows. In Section II, we introduce our proposed adaptive system including the process pipeline and principal transfer learning algorithm. Then, in Section III, a brief introduction of the dataset we used to conduct experiments and the procedure of experiments are given. Section IV presents the experiments results among our system and three state-of-art traditional machine learning methods. Section V and VI we present our findings, discussion and draw the conclusion, respectively.

II. METHODS

A. Transfer Component Analysis (TCA)

As illustrated in Section I, automatic pain assessment is a typical transductive transfer learning task where source domain D_S and target domain D_T are different while tasks are the same in both domains. A domain, as defined in [10], consists of two components, a feature space $\mathcal{X}$ and a marginal probability distribution *P(X)*, where *X* = $\{x_1,...,x_n\}$ $\in \mathcal{X}$. When feature spaces are the same in both domains, which is the case for automatic pain assessment study because physiological signal measurements remain the same, only the difference in marginal probability distribution makes it difficult to apply the model from the source domain to the target domain directly. Transfer Component Analysis (TCA) [13], thus, tries to map two domains into a new space and makes them share the same marginal probability distribution.

TCA assumes there exists a feature mapping ϕ such that $P(\phi(\mathbf{x}_s)) \approx P(\phi(\mathbf{x}_t))$ and thus, $P(y_s \mid \phi(\mathbf{x}_s))) \approx P(y_t \mid \phi(\mathbf{x}_t)))$ because we can assume that $P(y_s \mid \mathbf{x}_s)) \approx P(y_t \mid \mathbf{x}_t))$ which is only determined by the natural characteristic of our human body. In order to find such feature mapping ϕ, TCA utilizes a backward methods, which is to assume ϕ exists first and then use such ϕ to calculate the distance between domains and find the minimum value of distance. Thus, ϕ corresponding to the minimum distance is the desired mapping. The certain kind of distance TCA used is maximum mean discrepancy (MMD).

$$DISTANCE(\mathbf{x}_s, \mathbf{x}_t) = \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \phi(\mathbf{x}_i) - \frac{1}{n_2} \sum_{j=1}^{n_2} \phi(\mathbf{x}_j) \right\|_{\mathcal{H}}$$

Instead of finding the nonlinear transformation ϕ explicitly, TCA revisits a dimensionality reduction-based domain adaptation method. A kernel function $\mathbf{K}$ and a MMD matrix $\mathbf{L}$ are involved. Specifically, $\mathbf{K}_{s,s}$, $\mathbf{K}_{t,t}$, $\mathbf{K}_{s,t}$ represents the gram matrices defined on source domain, target domain and cross domain, respectively. Thus, the key problem transformed to solving the $\mathbf{K}$:

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{s,s} & \mathbf{K}_{s,t} \\ \mathbf{K}_{t,s} & \mathbf{K}_{t,t} \end{bmatrix} \quad (1)$$

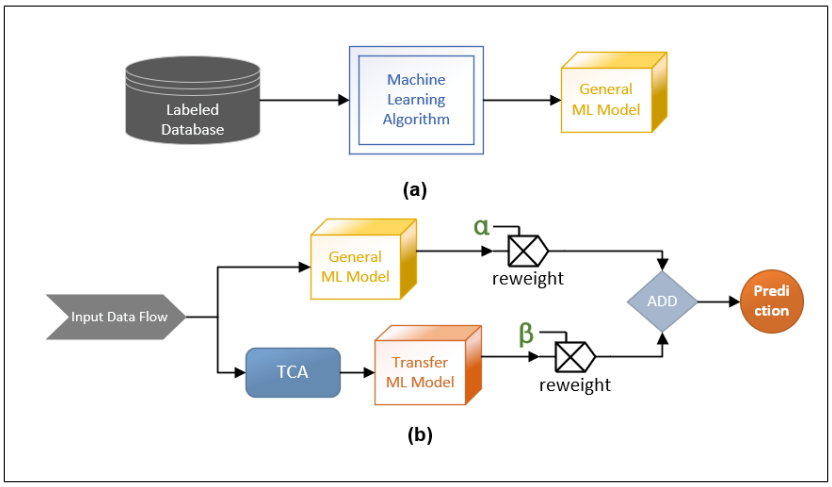


Fig. 1. ATLAS: Adaptive Transfer Learning Based Pain Assessment System

and the MMD matrix $\mathbf{L}$ is defined as:

$$\mathbf{L}_{ij} = \begin{cases} \frac{1}{{n_1}^2} & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_s, \\ \frac{1}{{n_2}^2} & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_t, \\ -\frac{1}{n_1 n_2} & \text{otherwise} \end{cases} \tag{2}$$

Therefore, the MMD equation can be transformed into:

$$\operatorname{tr}(\mathbf{KL}) - \lambda \operatorname{tr}(\mathbf{K}) \tag{3}$$

where $\operatorname{tr}(\cdot)$ represents trace of matrix. $-\lambda \operatorname{tr}(\mathbf{K})$ here is added manually and λ is a hyperparameter that the programmer can adjust. This term means when trying to minimize the distance between domains, TCA wants to remain the data variance, namely the divergence of the data, which is important for machine learning classification. Detailed explanation of equations can be found in Appendix. As of now, the problem is solvable but computation-consuming. Pan et al. [13] use a dimension-reduction methods to simplify the problem into:

$$\widetilde{\mathbf{K}} = \left(\mathbf{K}\mathbf{K}^{-1/2}\widetilde{\mathbf{W}}\right)\left(\widetilde{\mathbf{W}}^{\top}\mathbf{K}^{-1/2}\mathbf{K}\right) = \mathbf{K}\mathbf{W}\mathbf{W}^{\top}\mathbf{K} \tag{4}$$

and the optimization target becomes:

$$\begin{aligned} \min_{\mathbf{W}} \quad & \operatorname{tr}\left(\mathbf{W}^{\top}\mathbf{K}\mathbf{L}\mathbf{K}\mathbf{W}\right) + \mu \operatorname{tr}\left(\mathbf{W}^{\top}\mathbf{W}\right) \\ \text{s.t.} \quad & \mathbf{W}^{\top}\mathbf{K}\mathbf{H}\mathbf{K}\mathbf{W} = \mathbf{I}_m \end{aligned} \tag{5}$$

where $\mathbf{H} = \mathbf{I}_{n_1+n_2} - 1/\left(n_1+n_2\right)\mathbf{1}\mathbf{1}^{\top}$, $\mathbf{1} \in \mathbf{R}^{n_1+n_2}$ is the column vector with all 1's and W is the lower-dimensional intermediate matrix. The first m eigenvalues of W are the solution to the feature mapping ϕ.

Once the optimized feature mapping ϕ is solved, machine learning models can be trained in the new space $\{\phi(\mathbf{x}_s), y_s\}$ and applied on the mapped target data $\phi(\mathbf{x}_t)$.

B. Adaptive System

Traditional machine learning assumes test data share the same distribution as training data, while in many real-world scenarios, it doesn't hold. In the matter of pain assessment, individuals' demographic information (e.g., age, gender) and disease diagnosis may influence their physiological responses to pain and result in the different data distribution [14]. Hence, we propose an adaptive patient-centered automatic pain assessment system based on transfer component analysis algorithm as shown in Fig.1.

In the proposed system, a general ML (Machine learning) model is trained using a labeled database as shown in Fig.1 (a). Such labeled database can be a pre-experimented database on recruited volunteers or a public pain-related database such as BioVid [15] or X-ITE database [16]. After the general ML model is trained, as shown in Fig.1 (b), the new data flow will be input into two paths, the first path is the pre-trained general ML model and the second one is the TCA-based transfer ML model. In the second path, the labeled database and new data are mapped based on the TCA algorithm to minimize the difference in their distribution and then in the new mapping space, a transfer ML model is trained. Two classification results will be generated after two models and the results will be reweighted by adaptive hyperparameters α and β, ranging $[0, 1]$. These two hyperparameters are adapted by the confidence interval of the input data flow at the confidence level of 95% $\mathbf{c}$. The confidence interval shows the degree to which the actual value of this parameter has a certain probability of falling around the measurement result. It gives the degree of credibility of the measured value of the measured parameter. The confidence interval calculates as:

$$\mathbf{c} = \sqrt{\frac{Z^2 * p * (1-p)}{ss}} \tag{6}$$

where Z refers to Z value (e.g., 1.96 for 95% confidence level), p refers percentage picking a choice and ss refers to the number of samples. Note that each sample is a combination of feature values (i.e., one row in the dataframe) extracted from 5.5 second time window, instead of the discrete raw data point. The confidence interval shows whether the current samples can reflect the overall feature and such ability depends on the number of samples. In the beginning, α is set close to 1 and β is set to close 0 because there

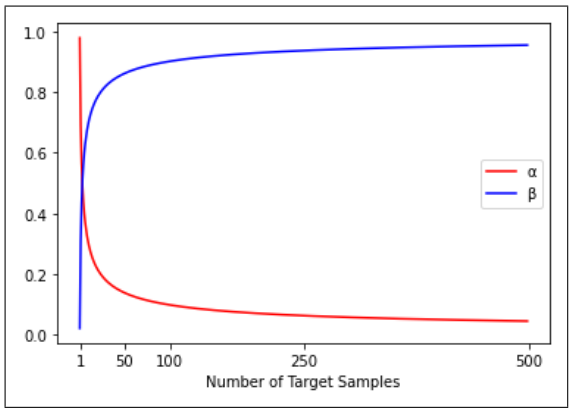


Fig. 2. The adaptive function for α and β

are no data for performing TCA, so the system entirely lies on the general ML model.As more input data comes, α adaptively drops and β adaptively increases following $\alpha = \mathbf{c}$ and $\beta = 1 - \mathbf{c}$ as shown in Fig. 2. The final result tends more to the general ML model at the beginning because there are not enough new samples to give a stable data distribution for transfer. Still, when new data samples are adequate to reflect the data distribution, the final result lies on the transfer ML model because the gap between the source domain and target domain is minimized. In the end, adding between the reweighted general ML model and the reweighted transfer ML model generates the final prediction result. The proposed system adaptively listens to general ML and transfer learning with the different input new samples to maintain high performance during real-life applications.

III. EXPERIMENTS

In this section, to evaluate our proposed adaptive system, three experiments are conducted on BioVid database comparing the proposed approach with references of Campbell et al. [18]'s working which used Linear Support Vector Machine (SVM), Gruss et al. [19]'s work which used SVM with radial basis function kernel (RBF kernel) and Lopez-Martinez et al. [20]'s work which used Neural Networks.

A. Database

The experiments are conducted on the BioVid database which is a public database that includes both biosignals and video data. It collected facial video, skin conductivity level (SCL), electrocardiogram (ECG) and sEMG (trapezius muscle, corrugator and zygomaticus) data with induced heat pain of four different pain levels on 87 healthy volunteers. Our study only focuses on physiological signals consisting of ECG, SCL, sEMG. Further detailed information on the database can be found in [15][17].

B. Reference Reproduction

As references, three existing works were reproduced on BioVid database, including the same data pre-processing techniques, feature extraction, classification model and evaluation metrics. Campbell et al. [18] collected the strengths of 13 existing feature extraction publications related to physiological signals and explored 155 different time domain and frequency domain features from ECG, SCL and sEMG signals. Then, a linear SVM classifier was implemented on 85 subjects from BioVid database and 90% accuracy was achieved. Gruss et al. [19] extracted a comprehensive and structured feature space containing 159 different features from ECG, SCL and sEMG. SVM with RBF kernel was chosen to classify baseline v.s. four different pain levels and results showed 79.29% - 90.94%. Lopez-Martinez et al. [20] extracted 12 features from SCL and five features from ECG, both in the time domain and utilized multi-task neural networks to classify binary tasks on four different pain levels on BioVid database. The results showed 82.75% accuracy in terms of baseline v.s. the highest pain level. As of data preprocessing, both [18] and [19] adopt Butterworth bandpass filter to denoise EMG and ECG signals with passbands of [20,250] Hz and [0.1,250] Hz, respectively. [20] only used [0.1,250] Hz Butterworth filter for ECG as EMG was not covered. Concerning QRS complex detection for R wave related features, [20] mentioned they utilized Pan-Tompkins algorithm [21] and the same QRS detection method was applied in our experiments.

The reference experiments consist of two parts: traditional ML reproduction and "Leave One Out" experiment. In traditional ML reproduction, the same data preprocessing steps, feature extraction and classifier model are deployed on the entire database with a 70:30 train-test split ratio. The data samples from all subjects are messed up and train test sets are split randomly. Then, ML evaluation metrics including accuracy, precision, recall, F1-score and area under the curve (AUC) of receiver operating characteristic (ROC) curve are calculated to assess the model. In the "Leave One Out" experiment, a similar ML pipeline is set, but the train-test split method changed to leave one subject as the test set. The rest 86 subjects as the training set and repeated the same experiment 87 times to loop all subjects and use the average evaluation score from 87 experiments as the final score. The "Leave One Out" algorithm is shown in Algorithm 1 and similar experiments are done in all three reference studies. In short, the "Leave One Out" algorithm sets one subject's data as the test set, and the rest 86 subjects' data as the training set and repeats the experiments 87 times.

Algorithm 1 Leave One Out Experiment

Input: 87 subjects' data of (X, y)
Output: evaluation metric scoring

```
scorings ← ∅
Data preprocessing
X' ← Feature extraction(X)
for each subject (X'_k, y_k) do
    clf_k ← trainMLModel(X̃'_k, ỹ_k)
    scorings ← clf_k.fit(X'_k, y_k)
end for
scoring ← average(scorings)
```

C. Adaptive system experiment

We conducted experiments to evaluate our proposed adaptive system compared to the reference studies. The adaptive system experiments follow the same framework as "leave one out" experiment that each subject will be set as the testing set and other subjects as the training set. However, for each subject, the new data comes in units of 10 data samples (i.e., 10 * 5.5s time window data) and at each unit, a new TL classifier is trained and reweighed according to the total of sample numbers as illustrated in Section II Part B.

Algorithm 2 Adaptive Transfer Learning Experiment

Input: 87 subjects' data of (X, y)
Output: evaluation metric scoring

$scorings \leftarrow \emptyset$
Data preprocessing
$X' \leftarrow$ Feature extraction(X)
for each subject (X'_k, y_k) **do**
  for numSample(i) from 0 to sizeof(X) **do**
    $(D_{t,i}, D_s) \leftarrow \text{TCA}(X'_{k,i}, \tilde{X}'_k)$
    $(y_{t,i}, y_s) \leftarrow (y_{k,i}, \tilde{y}_k)$
    $clf_{k,i} \leftarrow$ adaptiveTLSystem(D_s, y_s)
    $scorings \leftarrow clf_{k,i}$.fit$(D_{t,i}, y_{t,i})$
  end for
end for
$scoring \leftarrow$ average$(scorings)$

IV. RESULTS

Table I displays the classification accuracy from different experiments as illustrated in Section III under different tasks. As shown, traditional ML performs the best throughout all tasks and when it comes to "leave one out" experiment, the accuracy drops for 5.3, 6.6, 7.7, 8.1% in average for tasks of baseline v.s. pain level 1-4, respectively. Once our proposed adaptive transfer learning system is applied, the average accuracy increases for 4.3, 4.1, 5.2, 5.8% for tasks of baseline v.s. pain level 1-4, respectively. In addition, the average accuracy of the proposed system is 2.0% below the traditional supervised machine learning methods.

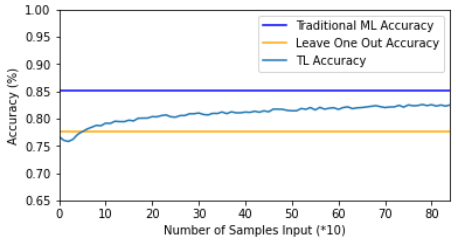
To compare the overall performance of models under three experiments, Fig. 3 shows the receiver operating characteristic (ROC) curve of each.

Fig. 4 shows the adaptive process of the proposed system. We hypothesize that the new samples have the same data distribution because the data are from a particular subject. The models for Traditional ML experiments and "Leave One Out" experiments remain unchanged. Thus, the prediction accuracy for Traditional ML experiments and "Leave One Out" experiments keep the same regardless of the number of new samples input. In Fig. 4, they are horizontal lines in blue and orange color. For the adaptive system experiments, the accuracy trend increases with the number of input samples increasing.

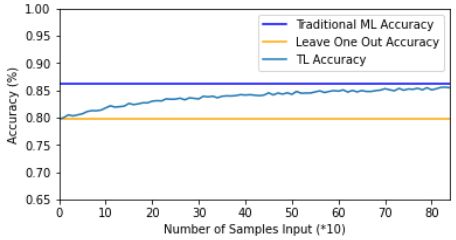
TABLE I
CLASSIFICATION ACCURACY FROM THREE EXPERIMENTS

Reference studies	Task	Trad[1]	LOO[2]	TL[3]
Campbell et al. [18]	BL0 v.s. PL1[4]	75.6	72.1	73.0
	BL0 v.s. PL2	78.3	**74.9**	73.6
	BL0 v.s. PL3	81.7	75.1	80.2
	BL0 v.s. PL4	85.1	77.6	82.3
Gruss et al. [19]	BL0 v.s. PL1	77.1	71.0	74.0
	BL0 v.s. PL2	78.1	72.1	76.8
	BL0 v.s. PL3	82.3	76.8	79.9
	BL0 v.s. PL4	86.2	79.7	**86.5**
Lopez et al. [20]	BL0 v.s. PL1	57.1	50.8	**59.7**
	BL0 v.s. PL2	63.4	54.9	**63.8**
	BL0 v.s. PL3	71.5	60.5	68.0
	BL0 v.s. PL4	80.8	69.3	75.3

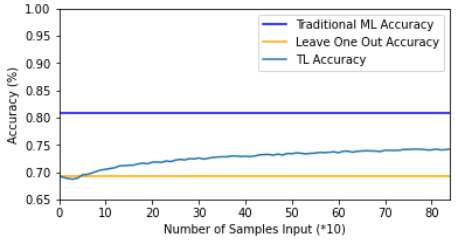
[1] Traditional Machine Learning Method, refer to Section III *Reference Reproduction Part 1*.
[2] "Leave One Out" Method, refer to Section III Algorithm 1.
[3] Adaptive Transfer Learning Method, refer to Section III Algorithm 2.
[4] BL0 refers to Baseline, PL1-4 refer to Pain level 1-4.



(a)

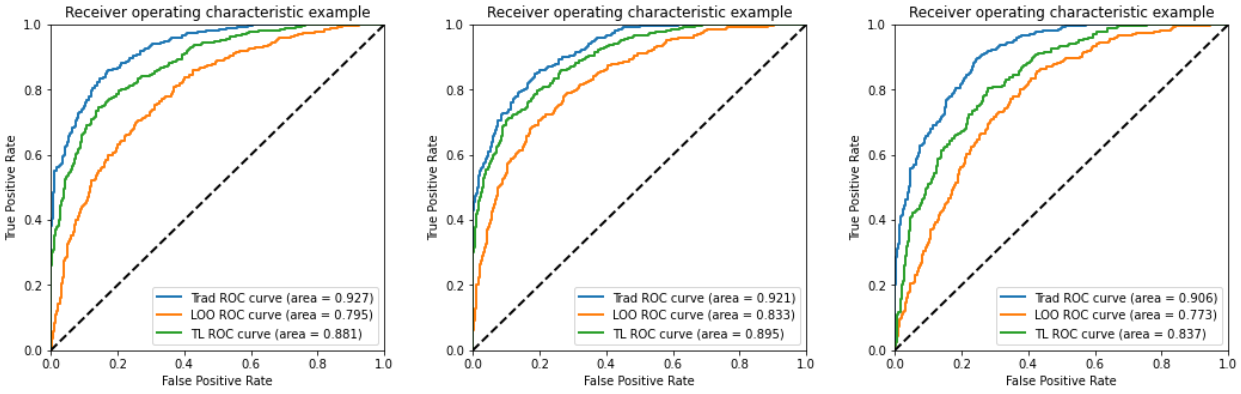


(b)



(c)

Fig. 4. Accuracy of traditional ML model, Leave One Out and our proposed adaptive transfer learning model ATLAS; based on (a) Campbell et al.[18], (b) Gruss et al. [19] and (c) Lopez et al. [20]



(a) (b) (c)

Fig. 3. ROC of traditional ML model, Leave One Out and our proposed adaptive transfer learning model ATLAS; based on (a) Campbell et al.[18], (b) Gruss et al. [19] and (c) Lopez et al. [20]

V. DISCUSSION

A. Performance Decline in Application Stage

As in Section I, current research primarily focuses on the verification stage in which they trained models on the entire dataset using random train-test split or cross-validation. However, due to the difference in data distribution, a general model may not fit on a particular subject. Our study tends to verify this hypothesis and the results showed a 5.5 to 8.1% accuracy reduction between the traditional ML experiments and the "Leave One Out" experiments in which we mimic the scenarios in real life. Similar results can also be found in the ROC plots. Further studies can aim to put the experiments on the applications scenarios and solve the problem of performance decline.

B. Evaluation of the Proposed System

Experiments were conducted comparing three existing works to evaluate the proposed adaptive transfer learning system. Accuracy and area under the curve (AUC) of the ROC curve are calculated as the metrics to evaluate models.

The experiments results show an overall trend that *Traditional ML* $>$ *Transfer System* $>$ *Leave One Out* in classification performance. The average accuracy increases for 4.3, 4.1, 5.2 and 5.8% for tasks of baseline v.s. pain level 1-4 by using our proposed adaptive system, which verifies the system's efficacy. In addition, the average gap between our proposed methods and the traditional methods is 2% in accuracy, which is a good number. Notice that the traditional machine learning methods use labeled testing sets while our system uses unlabeled test sets. Thus, it is encouraging to see as small as a 2% difference. Four results out of all experiment results don't support the inequality. They are either *Traditional ML* $<$ *Transfer System* or *Transfer System* $<$ *Leave One Out*. We think the reason may lie in machine learning models' random state. For lower pain level experiments, since the models don't have high confidence and performance as models for high pain level tasks, the effect of random states becomes more powerful and the classification accuracy, thus, is easily shaken by the random states.

In the adaptive process of the proposed system, as shown in Fig. 4, the classification accuracy has the trend of increasing and tends to have an upper limit as the number of input samples increases. This phenomenon is in line with our expectations. Nevertheless, there are sharp declines in the experiments with Campbell et al. and Lopez et al.'s work. We think this is because, at the beginning of the experiments, the information provided by the newly input data is not adequate compared to the weights put on its side. In other words, the input data are not sufficient to represent a data distribution for the transfer learning algorithm so that the transfer model doesn't perform well. Still, weights are already settled on the transfer side and therefore, the overall system performance drops. To solve this problem, future studies can focus on detailed optimization of the adaptive process to let the system have better performance regardless of the number of input samples. For example, at the beginning of X samples, the weights are set as $\alpha = 1, \beta = 0$ to remove the drop.

As experiments embodied, the proposed ATLAS system is a "framework" rather than a particular algorithm. The reason behind it is that the system only works on the instances, i.e., data themselves independent with the classification model. Thus, various machine learning algorithms can be embedded in the proposed system, e.g., SVM, neural networks, which makes the system flexible and easy to be extended to future studies. Deep Learning algorithms show promising results in terms of physiological signal analysis in some cases but due to the limited sample size of BioVid and deep learning studies, we are only considering traditional machine learning algorithms. However, we believe that deep learning can also be embedded with the proposed ATLAS system as only data is involved in the proposed system.

VI. CONCLUSION

In this paper we propose ATLAS, a novel unsupervised adaptive transfer learning system for automatic pain assessment. ATLAS uses transfer component analysis, an effective

transfer learning algorithm as the fundamental learning algorithm, and adopts statistical theory to adapt the transfer learning component and traditional machine learning component to achieve an overall better performance. The primary purpose of this study is to compensate for the deficiency of a real-life oriented automatic pain assessment study. Three different experiments were conducted including 1) a traditional machine learning study, in which we reproduce three existing works, 2) a "Leave One Out" experiment, in which we tested the performance of the three existing works under real-life scenario and 3) an adaptive system experiments, in which we evaluated our proposed adaptive system. The experiments results showed that when state-of-art methods were used in real-life scenarios, the classification accuracy dropped by 8.1%. ATLAS compensates the drop and increase the classification accuracy by 6%. More importantly, our proposed unsupervised system with unlabeled data achieves high accuracy as close as just 2% below the accuracy of supervised machine learning models with labeled data.

REFERENCES

[1] Yong, R. Jasona,*; Mullins, Peter M.b; Bhattacharyya, Neilc Prevalence of chronic pain among adults in the United States, PAIN: April 02, 2021 - Volume - Issue - doi: 10.1097/j.pain.0000000000002291

[2] Paice, Judith A., and Jamie H. Von Roenn. "Under-or overtreatment of pain in the patient with cancer: how to achieve proper balance." Journal of Clinical Oncology 32, no. 16 (2014): 1721-1726.

[3] Pouromran, Fatemeh, Srinivasan Radhakrishnan, and Sagar Kamarthi. "Exploration of physiological sensors, features, and machine learning models for pain intensity estimation." Plos one 16, no. 7 (2021): e0254108.

[4] Huang, Yibo, Linbo Qing, Shengyu Xu, Lu Wang, and Yonghong Peng. "HybNet: a hybrid network structure for pain intensity estimation." The Visual Computer (2021): 1-12.

[5] Werner, Philipp, 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).

[6] Kächele, Markus, Patrick Thiam, Mohammadreza Amirian, Friedhelm Schwenker, and Günther Palm. "Methods for person-centered continuous pain intensity assessment from bio-physiological channels." IEEE Journal of Selected Topics in Signal Processing 10, no. 5 (2016): 854-864.

[7] Kächele, Markus, 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, no. 1 (2017): 71-83.

[8] Chen, Jixu, Xiaoming Liu, Peter Tu, and Amy Aragones. "Learning person-specific models for facial expression and action unit recognition." Pattern Recognition Letters 34, no. 15 (2013): 1964-1970.

[9] Zhuang, Fuzhen, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. "A comprehensive survey on transfer learning." Proceedings of the IEEE 109, no. 1 (2020): 43-76.

[10] Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." IEEE Transactions on knowledge and data engineering 22, no. 10 (2009): 1345-1359.

[11] Huang, Jiayuan, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. "Correcting sample selection bias by unlabeled data." Advances in neural information processing systems 19 (2006): 601-608.

[12] Dai, Wenyuan, Gui-Rong Xue, Qiang Yang, and Yong Yu. "Transferring naive bayes classifiers for text classification." In AAAI, vol. 7, pp. 540-545. 2007.

[13] Pan, Sinno Jialin, Ivor W. Tsang, James T. Kwok, and Qiang Yang. "Domain adaptation via transfer component analysis." IEEE transactions on neural networks 22, no. 2 (2010): 199-210.

[14] Fang, Ruijie, 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 International Conference on Intelligent Medicine and Image Processing (IMIP).

[15] Walter, Steffen, 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), pp. 128-131. IEEE, 2013.

[16] Gruss, Sascha, Mattis Geiger, Philipp Werner, Oliver Wilhelm, Harald C. Traue, Ayoub Al-Hamadi, and Steffen Walter. "Multi-modal signals for analyzing pain responses to thermal and electrical stimuli." JoVE (Journal of Visualized Experiments) 146 (2019): e59057.

[17] Werner, Philipp, Ayoub Al-Hamadi, Robert Niese, Steffen Walter, Sascha Gruss, and Harald C. Traue. "Towards pain monitoring: Facial expression, head pose, a new database, an automatic system and remaining challenges." In Proceedings of the British Machine Vision Conference, pp. 1-13. 2013.

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

[19] Gruss, Sascha, 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, no. 10 (2015): e0140330.

[20] Lopez-Martinez, Daniel, 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), pp. 181-184. IEEE, 2017.

[21] Pan, Jiapu, and Willis J. Tompkins. "A real-time QRS detection algorithm." IEEE transactions on biomedical engineering 3 (1985): 230-236.