

# Student Engagement from Video using Unsupervised Domain Adaptation

Chinchu Thomas<sup>a</sup>, Seethamraju Purvaj and Dinesh Babu Jayagopi

Multimodal Perception Lab, International Institute of Information Technology Bangalore (IIIT-B), Karnataka, India

**Keywords:** Student Engagement, Unsupervised Domain Adaptation, Discrepancy and Adversarial Methods.

**Abstract:** Student engagement is the key to successful learning. Measuring student engagement is of utmost importance in the current global scenario where learning happens over online platforms. Automatic analysis of student engagement, in offline and online social interactions, is largely carried out using supervised machine learning techniques. Recent advances in deep learning have improved performance, albeit at the cost of collecting a large volume of labeled data, which can be tedious and expensive. Unsupervised domain adaptation using the deep learning technique is an emerging and promising direction in machine learning when labeled data is less or absent. Motivated by this, we pose our research question: "Can deep unsupervised domain adaptation techniques be used to infer student engagement in classroom videos with unlabeled data?" In our work, two such classic techniques i.e. Joint Adaptation Network and adversarial domain adaptation using Wasserstein distance were explored for this task and posed as a binary classification problem along with different base models such as ResNet and I3D. The best-obtained result using the JAN network has an accuracy of 68% and f1-score of 0.80 for binary student engagement with RGB-I3D network as the base model. The adversarial domain adaptation method gave an accuracy of 71% and f1-score of 0.82 with ResNet 50 as the feature extractor for predicting the engagement of the students in the classroom.

## 1 INTRODUCTION


Student engagement is essential to successful learning and refers to the extent to which students are interested, attentive, and curious when they are learning or being taught. In the current global scenario, where online education has become inevitable, it is of utmost importance to track student or user engagement. Automatic analysis of student engagement has been previously done using supervised machine learning methods (Nezami et al., 2019) (Thomas and Jayagopi, 2017) (Whitehill et al., 2014).

Recent advances in deep learning have improved performance, but come at an increased cost of collecting large volumes of labeled data for training the network. At the same time, there is already an ample amount of annotated data available for various domains and tasks. Unsupervised domain adaptation (UDA) using deep neural networks is an emerging and promising trend in machine learning to utilize the existing labeled data. Domain adaptation aims to learn a concept from labeled data in a source domain that performs well on a different but related target domain that can be labeled, partially labeled, or

unlabeled. Unsupervised domain adaptation specifically addresses the scenario where the source data is labeled, and the target data is unlabeled.

In this work, we address the following research question: Can deep unsupervised domain adaptation techniques be used to infer student engagement from classroom videos with unlabeled data? We address this question by using a deep learning pipeline that embeds unsupervised domain adaptation into it. To attempt this, we experimented with two methods: Joint Adaptation Network (JAN) (Long et al., 2017) and an adversarial method, that uses Wasserstein distance for unsupervised domain adaptation (Drossos et al., 2019). Along with the UDA, we experimented with different base models in the deep learning pipeline to understand the impact of each model on the domain adaptation methods. Also, we explored the effectiveness of image-based models and video-based models for unsupervised domain adaptation by using various pre-trained models like ResNet 18, ResNet 50, and RGB-I3D.

The contributions of the paper are as follows: We propose an unsupervised method for inferring the engagement level of the students in a classroom, unlike previous works which use supervised learning meth-

<sup>a</sup>  <https://orcid.org/0000-0003-4887-2273>

ods. Although the joint adaptation network and adversarial adaptation using Wasserstein distance were originally proposed for images, in this work, we use it on video data. We experimented with three different base models; ResNet 18 pre-trained on ImageNet data (He et al., 2016), ResNet 50 pre-trained on VGFace2 (Cao et al., 2018) and RGB-I3D pre-trained on ImageNet and Kinetics dataset (Carreira and Zisserman, 2017). Also, we did a comparative study of supervised and unsupervised learning methods to see how far the unsupervised domain adaptation can perform.

The rest of the paper is organized as follows. Section 2 discusses the related works relevant to the automatic prediction of student engagement and unsupervised domain adaptation. The theoretical details of the methods that we used for the experiments are described in Section 3. The experimental details including the dataset, implementation details and the baselines are outlined in Section 4. Section 5 presents the results and the findings. Section 6 concludes the paper.

## 2 RELATED WORK

In this work, we focus on inferring student engagement using unsupervised domain adaptation. Hence, this section discusses works related to the task as well as the modeling approach.

**Task: Student Engagement Analysis.** Student engagement can be predicted in different ways such as from a camera, logs from the learning platforms, and data from wearable sensors. In this work, we focus on the video data captured from an RGB camera. The different learning environments can be a classroom (Thomas and Jayagopi, 2017) (Raca, 2015), online learning platforms (Gupta et al., 2016) (Grafsgaard et al., 2013) or intelligent tutoring systems (Whitehill et al., 2014). In all these settings, the goal is to infer the affective state of engagement using supervised learning techniques. These works utilized either traditional machine learning methods or a deep learning pipeline to infer the engagement level of the students.

Supervised methods used handcrafted features such as eye gaze, head pose, and facial action unit intensities to capture the engagement state. Raca (Raca, 2015) used features computed from motion detection, head detection, and estimation of orientation. Whitehill et al. (Whitehill et al., 2014) worked on facial expressions to predict different levels of engagement in an interactive learning environment. Thomas and Jayagopi (Thomas and Jayagopi, 2017) utilized the

head pose, eye gaze, and facial action unit intensities to infer the state of students.

Several works utilized deep learning models such as Inception model, C3D, and LRCN networks on frame level as well as video level to infer the affective states of engagement, boredom, confusion and frustration on DaiSEE dataset (Gupta et al., 2019) (Gupta et al., 2016) (Ashwin and Guddeti, 2019). Gupta et al. Yang et al. (Yang et al., 2018) used a multi-modal regression model based on the multi-instance mechanism as well as LSTM to predict the engagement intensity for the engagement in the wild dataset from the EmotiW challenge. Thomas et al. (Thomas et al., 2018) used Temporal Convolutional Networks (TCNs) to predict engagement intensity on engagement in the wild dataset. More recently, a several attempts have been made to predict student engagement in online learning (Abedi and Khan, 2021b) (Abedi and Khan, 2021a) (Geng et al., 2019) (Huang et al., 2019) (Liao et al., 2021) (Wang et al., 2020) (Zhang et al., 2019).

### **Modeling: Unsupervised Domain Adaptation.**

Deep domain adaptation (DA) has emerged as a new learning paradigm to address the lack of huge amounts of labeled data. The conventional methods learn a shared feature subspace or reuse general source examples with shallow representations, whereas deep domain adaptation methods make use of deep networks to learn more transferable representations by embedding domain adaptation in the pipeline of deep learning. Wang et al. (Wang and Deng, 2018) described the different methods such as discrepancy based, adversarial-based and reconstruction-based deep DA approaches.

In this work, we use a discrepancy-based DA method (JAN network described in Section 3.1) which uses maximum mean discrepancy as the statistic criterion to learn the invariant features. The adversarial method (adversarial network described in Section 3.2) is a non-generative model which learns domain invariant representations. These representations are learned with a feature extractor that learns a discriminative representation using the labels in the source domain and maps the target data to the same space through a domain-confusion loss. There are applications of unsupervised domain adaptation for vision applications, NLP tasks, and time-series data. A detailed study on the theoretical aspects and the applications are done in (Wang and Deng, 2018) (Wilson and Cook, 2020). In this work, we focus on the problem of inferring student engagement. The models were trained to learn invariant features that can be transferred from an online learning environment to a classroom setting.

### 3 UNSUPERVISED DOMAIN ADAPTATION USING JAN AND ADVERSARIAL APPROACHES

In this section, we briefly explain unsupervised domain adaptation using joint adaptation network (JAN) as well as adversarial approach.

#### 3.1 Joint Adaptation Network

In unsupervised domain adaptation, we are given a source domain  $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$  of  $n_s$  labeled examples and a target domain  $\mathcal{D}_t = \{(x_i^t, y_i^t)\}_{i=1}^{n_t}$  of  $n_t$  unlabeled examples. The source and target domains are sampled from joint distributions  $P(X^s, Y^s)$  and  $Q(X^t, Y^t)$  respectively, where  $P \neq Q$ .

##### 3.1.1 Joint Adaptation Network

The underlying idea in Joint Adaptation Network (JAN) (Long et al., 2017) is to extend deep convolutional neural networks (CNNs) with additional fully connected layers to learn a joint representation of the data and the label. CNNs are known to learn generic features in the convolutional layers and domain-specific features in the final layers. The convolutional features are transferable across the domains, while the features in the fully connected layers cannot be transferred safely for domain adaptation due to cross-domain discrepancy. Additionally, the shift in the labels lingers in the classifier layers. In unsupervised domain adaptation, the joint distribution of the features in the higher layers are matched for source and target domain using joint maximum mean discrepancy (JMMD).  $\mathcal{L}$  denotes the domain-specific layers where the activations are not safely transferable. By integrating the JMMD over the domain-specific layers  $\mathcal{L}$  into the CNN error, the joint distributions are matched end-to-end with network training,

$$\min_f \frac{1}{n_s} \sum_{i=1}^{n_s} J(f(x_i^s), y_i^s) + \lambda \hat{D}_{\mathcal{L}}(P, Q) \quad (1)$$

where  $J(\cdot)$  is the cross entropy loss,  $\hat{D}_{\mathcal{L}}(P, Q)$  is the JMMD penalty and  $\lambda > 0$  is a trade-off parameter of the JMMD penalty.

#### 3.2 Adversarial Domain Adaptation using Wasserstein Distance

Let  $\mathcal{D}_S = \langle \mathcal{Z}_S, f_S \rangle$  and  $\mathcal{D}_T = \langle \mathcal{Z}_T, f_T \rangle$  be the source and target domains, respectively. Let  $h$  be the classifier,  $z \sim \mathcal{Z}$  is the input to the labeling process  $f$ . The

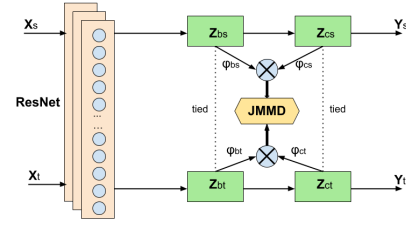


Figure 1: Joint Adaptation Network.

aim of unsupervised DA is to obtain a classifier  $h$  that yields a low error in the source domain and adapt it to get a low target classification error without using labels from the target domain during the adaptation process.

The whole process happens in two stages. The first stage is pre-training, where a feature extractor  $M_S$  is obtained during the optimization of the label classification, and the second stage is the adaptation process, where a copy  $M_T$  of  $M_S$  is further optimized during the adversarial training.

In this framework, Drossos et al. (Drossos et al., 2019) employ a deep neural network (DNN) and the Wasserstein generative adversarial networks (WGAN) formulation and algorithm (Arjovsky et al., 2017). DNN consists of a feature extractor  $M$ , a label classifier  $h$ , and a domain classifier  $h_d$ . There are two steps involved in the formulation.

The first step (pre-training) is to optimize  $M$  and  $h$  using the labeled data  $(x_s, y_s)$  from source domain  $(\mathbb{X}_S, \mathbb{Y}_S)$ , where  $y_S$  is 1-hot encoding of the classes and the binary cross-entropy as the loss function  $\mathcal{E}_S(h, f_S)$  in source domain:

$$\mathcal{L}_{labels}(h, M) = - \sum_{(x,y) \in (\mathbb{X}_S, \mathbb{Y}_S)} y^T \log(h(M(x))) \quad (2)$$

and the classifier  $h^*$  is obtained and the source domain feature extractor  $M_S$  by

$$h^*, M_S = \arg \min_{h, M} \mathcal{L}_{labels}(h, M) \quad (3)$$

$w_{M_S}$  denotes the parameters of the feature extractor  $M_S$  and will be used as initial values for the adapted feature extractor  $M_T$ .

In the second step (adaptation),  $M_S$  is adapted to the target domain using an adversarial training procedure. In this framework, order-1 Wasserstein distance is used as the metric to measure the discrepancy between  $\mathcal{Z}_S$   $\mathcal{Z}_T$ . The process of the adaptation of  $M_T$  is performed by the iterative minimization of the losses,

$$\sum_{x \in \mathbb{X}_S} h_d(M_S(x)) - \sum_{x \in \mathbb{X}_T} h_d(M_T(x)), \quad (4)$$

$$\sum_{x \in \mathbb{X}_T} h_d(M_T(x)) + \mathcal{L}_{labels}(h^*, M_T). \quad (5)$$

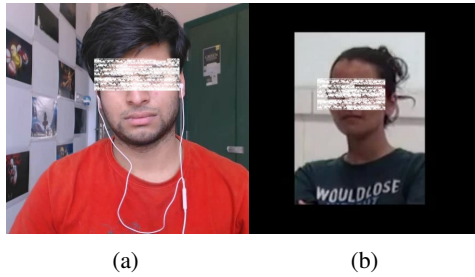


Figure 2: Sample frames from (a) DaiSEE dataset and (b) SEC dataset.

## 4 EXPERIMENTS

In this section, we discuss the details of the datasets, baseline models and computational descriptors used in the experiments.

### 4.1 Datasets

The details of the source and target datasets are explained in the following paragraphs.

**Source Dataset.** The source dataset used for the experiments is the in-the-wild DaiSEE dataset consisting of 9068 videos of 112 users annotated for affective states of boredom, engagement, confusion, and frustration (Gupta et al., 2016). The data comprises videos of users in an e-learning environment. The affective states are annotated for four levels: very-low, low, high, and very-high from the CrowdFlower platform. The videos are of 10-second duration. All the videos are annotated for multiple affect since a person may show different effects while in the learning environment. This work considers only the engagement affective state. The videos in very low and low levels are combined for the distracted label (label '0') and high and very-high are combined for the engaged label (label '1').

**Target Dataset.** The target dataset is the X dataset which consists of 2262 videos of students attending video presentations projected on the screen in a classroom. There are 10 unique students in the classroom, and each student video is trimmed for 10 seconds. The student affective state of engagement is annotated by external observers on a binary scale of engaged or distracted. More details of the dataset can be found in (Thomas and Jayagopi, 2017). The dataset is highly imbalanced. For the experiments, the train, validation, and test set consist of 1064, 456, and 742 samples respectively. Sample frames from the videos are shown in Fig. 2.

### 4.2 Modeling

In this subsection, we describe our baseline models and the set of computational descriptors for further modeling.

**Baseline:** The baselines are created for the target test set (IITB-SE) to compare how well the unsupervised domain adaptation model performs compared to the supervised and unsupervised learning task. The baseline for the supervised learning is the majority baseline which is created for the test set. For the supervised method, we considered Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) classifiers. For the unsupervised domain adaptation, the fine-tuned source models before adaptation were considered as baselines. We considered ResNet 18 pre-trained on ImageNet; ResNet 50 pre-trained on VGGFace2 dataset and RGB-I3D model pre-trained on ImageNet and Kinetics dataset as base models in JAN framework and as generators in an adversarial setting.

**Computational Descriptors:** The models, with supervised learning, used visual features. The visual features were computed using OpenFace toolbox (Baltrušaitis et al., 2016) for every frame. We extracted features related to eye gaze, head pose, and facial action unit (AUCs) of the speaker in the video. The mean and standard deviation of the features were computed to aggregate the statistics to a video level, which resulted in a 46-dimensional feature vector. We ran feature selection to choose 39 features that resulted in the highest relevance score. Another set of visual features that we used for the experiment was the 512-dimensional feature vector extracted from the videos. These are computed from the last pooling layer of the ResNet 18 pre-trained network. The third set of features used for the experiments were the last layer 2048-dimensional features from ResNet 50 VGGFace2 pre-trained network. Also, we computed features from the last layer of the RGB-I3D model pre-trained on ImageNet. The feature vectors from both ResNet 18 and ResNet 50 were computed for the frames, and the mean of all the frames was considered as the final feature vector.

### 4.3 Implementation Details

#### 4.3.1 JAN Network

The details of the architecture are described in this section. The base network of the model were pre-trained ResNet models and I3D model. The exper-



iments were done using ResNet 18, ResNet 50, and I3D architectures. The ResNet models until the last pooling layer were used as the base network. The ResNet 18 base network is followed by a bottleneck layer and a classifier layer. The bottleneck layer is a fully-connected layer with  $512 \times 256$  neurons and the classifier layer is also a fully connected layer with  $256 \times 2$  neurons. The ResNet 50 base network is followed by the bottleneck layer with  $2048 \times 256$  neurons and the classifier layer with  $256 \times 2$  neurons. The RGB-I3D base model is followed by a bottleneck layer with  $2048 \times 256$  neurons and a classifier layer with  $256 \times 2$  neurons. The output from the classifier layer is passed through a Softmax layer for the class probabilities. The bottleneck layer is followed by ReLU activation and drop out of 0.2. The bottleneck layer and the classifier layers were initialized at random with a normal distribution. We fine-tune all convolutional and pooling layers and train the classifier layer via backpropagation. Since the classifier is trained from scratch, we set its learning rate to be 10 times that of the other layers and the initial learning rate was 0.01. We use mini-batch stochastic gradient descent (SGD) with a momentum of 0.5 and the learning rate annealing strategy in RevGrad (Long et al., 2017). The trade-off parameter of the JMMD term was 0.5. We used the PyTorch implementation of JAN for the experiments<sup>1</sup>. The learning rate is not selected by a grid search due to high computational cost; it is adjusted during SGD as in the original implementation of JAN.

The training process is as follows: The videos are sampled at 1 frame per second to extract the frames of the videos. The frames are cropped for getting the frontal faces using dlib library, removing all unnecessary background information. We did this preprocessing with the assumption that student engagement can be learned from the region above the shoulder. The video frames are resized to  $456 \times 256$ . During training, a random  $224 \times 224$  pixels spatial crop of a random frame of the visual data is randomly flipped in the left/right direction and fed into the model. The activities of the penultimate layer are spatially pooled. This output from the base network is passed to the bottleneck layer. The network is trained for 20000 iterations. Over multiple iterations, we assume that the model is able to learn the temporal dependencies in the video. The validation is done at every 500 iterations inside the training using the above-mentioned procedure. The models which resulted in the best validation accuracy are tested. During testing, the entire video data is fed into the network one frame at a time. The network predicts for every frame in a video,

and the final prediction is the majority label computed over all the frames. For the RGB-I3D model, the entire video which is sampled at 1 frame per second is passed to the model directly for training and testing.

### 4.3.2 Adversarial Network

The architectural details of the generator models in the adversarial method are explained in this section. The first network that we used as the generator, was ResNet 18 pre-trained on ImageNet. The second network considered was pre-trained ResNet 50 as  $M_S$ . ResNet 50 is pre-trained on the VGGFace2 dataset, which is closer to the source and target dataset. The last one was RGB-I3D pre-trained on ImageNet and Kinetics dataset. The discriminator has two fully connected layers followed by a ReLU. The classifier layer consists of three feed-forward layers, each one followed by a ReLU and a drop out of 0.25. The feed-forward layers consists of 512, 256, 2 neurons for ResNet 18 model; 2048, 256, 2 neurons for ResNet 50 model and 2048, 256, 2 neurons for RGB-I3D model. The output non-linearity of  $h$  is the softmax function. The experiments were done using the PyTorch<sup>2</sup> implementation of the adversarial model. For the adaptation stage, the RMSProp optimizer is used with a learning rate of  $5 \times 10^{-4}$  for RGB-I3D and  $1 \times 10^{-3}$  and  $5 \times 10^{-4}$  for the generator and discriminator respectively in ResNet based models.

The training process for the model is as follows: The videos are sampled at 1 frame per second to extract the frames of the videos. The frames are cropped for getting the frontal faces. A random frame is sampled, and the frame is resized to  $224 \times 224$ . In the first step, the ResNet 18 model ( $M_S$ ) pre-trained on ImageNet, and  $h$  are trained using the source dataset in a supervised manner. This resulted in an adapted feature extractor  $M_S$  on the source dataset with an accuracy of 0.88. The ResNet 50 model resulted in an accuracy of 63% while pre-training. The RGB-I3D model resulted in an accuracy of 67%. In the second step of the adaptation,  $M_T$  is initialized with adapted  $M_S$ . The network is trained to fool the domain classifier. The adaptation process is done for 300 epochs and all the models are tested. The results are reported for the best model. During testing, the video data are fed into the model one frame at a time. The testing is done using the adapted feature extractor and the label classifier. The network predicts for every frame in a video, and the final prediction is the majority label computed over all the frames. For I3D, the entire video is passed for testing the models.

<sup>1</sup><https://github.com/thuml/Xlearn>

<sup>2</sup><https://github.com/dr-costas/undaw>

Table 1: Student engagement classification result with supervised learning.

| Model               | Feature     | Accuracy | Precision | Recall | f1-score |
|---------------------|-------------|----------|-----------|--------|----------|
| Majority Baseline   |             | 0.83     | 0.83      | 1.0    | 0.91     |
| Naive Bayes         | facial cues | 0.16     | 0.0       | 0.0    | 0.0      |
| Logistic Regression | facial cues | 0.83     | 0.83      | 1.0    | 0.91     |
| SVM                 | facial cues | 0.83     | 0.83      | 1.0    | 0.91     |
| Random Forest       | facial cues | 0.75     | 0.82      | 0.90   | 0.86     |
| Naive Bayes         | ResNet 18   | 0.72     | 0.89      | 0.76   | 0.82     |
| Logistic Regression | ResNet 18   | 0.84     | 0.90      | 0.91   | 0.90     |
| SVM                 | ResNet 18   | 0.82     | 0.91      | 0.86   | 0.89     |
| Random Forest       | ResNet 18   | 0.86     | 0.88      | 0.95   | 0.92     |
| Naive Bayes         | ResNet 50   | 0.78     | 0.89      | 0.85   | 0.87     |
| Logistic Regression | ResNet 50   | 0.83     | 0.86      | 0.95   | 0.90     |
| SVM                 | ResNet 50   | 0.83     | 0.88      | 0.92   | 0.90     |
| Random Forest       | ResNet 50   | 0.85     | 0.88      | 0.96   | 0.91     |
| Naive Bayes         | RGB-I3D     | 0.80     | 0.90      | 0.86   | 0.88     |
| Logistic Regression | RGB-I3D     | 0.81     | 0.90      | 0.88   | 0.89     |
| SVM                 | RGB-I3D     | 0.81     | 0.90      | 0.86   | 0.88     |
| Random Forest       | RGB-I3D     | 0.85     | 0.87      | 0.96   | 0.91     |

Table 2: Student engagement classification result with source domain fine-tuning (before adaptation).

| Base model          | Accuracy | Precision | Recall | f1-score |
|---------------------|----------|-----------|--------|----------|
| ResNet18 (ImageNet) | 0.19     | 0.79      | 0.04   | 0.11     |
| ResNet50 (VGGFace2) | 0.21     | 0.90      | 0.06   | 0.11     |
| RGB-I3D (ImageNet)  | 0.41     | 0.77      | 0.40   | 0.53     |

Table 3: Student engagement classification result with unsupervised domain adaptation.

| UDA method  | Base model           | Accuracy    | Precision | Recall | f1-score    |
|-------------|----------------------|-------------|-----------|--------|-------------|
| JAN         | ResNet18 (ImageNet)  | 0.61        | 0.82      | 0.74   | 0.78        |
| JAN         | ResNet50 (VGGFace2)  | 0.57        | 0.81      | 0.63   | 0.71        |
| JAN         | RGB-I3D (ImageNet)   | 0.68        | 0.81      | 0.80   | 0.80        |
| Adversarial | ResNet18 (ImageNet)  | 0.62        | 0.88      | 0.62   | 0.73        |
| Adversarial | ResNet 50 (VGGFace2) | <b>0.71</b> | 0.84      | 0.80   | <b>0.82</b> |
| Adversarial | RGB-I3D (ImageNet)   | 0.54        | 0.83      | 0.59   | 0.69        |

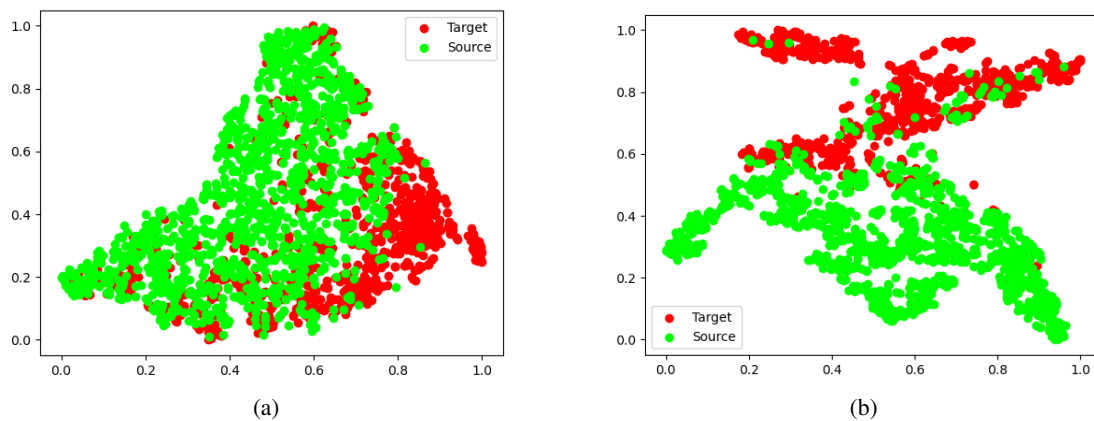


Figure 3: t-SNE plots for adversarial model with ResNet 50 as the base model (a) before adaptation (b) after adaptation.

## 5 RESULTS AND DISCUSSION

In this section, we describe the results for the binary classification task of engaged or distracted. The metrics used for evaluation are accuracy, precision, recall, and f1-score. The results are reported in Table 1. The majority classifier, which classifies everything to the majority class, resulted in a test accuracy of 83% and an f1-score of 0.68 on the SE test data. Supervised learning methods such as Logistic Regression and SVM performed just like a majority classifier. Naive Bayes was putting everything to the minority class, which is even worse. Random Forest performed better on facial cues with 75% accuracy and f1-score of 0.86. When the features were extracted from the last layer of ResNet 18 pre-trained on ImageNet, there is an improvement in the performance, especially for Naive Bayes and Random Forest classifiers. Naive Bayes performed with an accuracy of 72%, f1-score of 0.82, and Random Forest with 86% accuracy and 0.92 as f1-score. When the features were extracted from ResNet 50 pre-trained on VGGFace2, there is a slight drop in the performance compared to ResNet 18 pre-trained on ImageNet. This can be because of the huge dimensionality of the feature vector compared to the number of samples in the training set. This observation is consistent with the RGB-I3D model as well. The results from the supervised learning methods reveal the difficulty with respect to the dataset, and the low performance also stems from the huge class imbalance in the SE dataset.

In Table 2, we can see the results of the base models fine-tuned with the source dataset and then tested on the target dataset. This is considered as the baseline for the UDA methods. The results show that only the RGB-I3D model is able to perform a little better, with an accuracy of 41% and f1-score of 0.53 on a similar dataset. These results show the requirement to adapt the models for a specific domain of the dataset.

The results for unsupervised domain adaptation are reported in Table 3. In unsupervised domain adaptation, we experimented with discrepancy-based (JAN) and adversarial-based domain adaptation methods coupled with various pre-trained deep networks. In both the cases, the results are reported ResNet 18 pre-trained on ImageNet, ResNet 50 pre-trained on VGGFace2 dataset, and RGB-I3D pre-trained on ImageNet and Kinetics datasets. Among the JAN models, the one with RGB-I3D pre-trained on ImageNet and Kinetics datasets as the base network performed better than other base models, with an accuracy of 68% and f1-score of 0.80. Among the adversarial method, the base model with ResNet 50 pre-trained on VGGFace2 performed way better than all other

models, with an accuracy of 71% and f1-score of 0.82. This model performed significantly better than all the discussed unsupervised models. The t-SNE plot for the best model reported, ResNet 50 pre-trained with VGGFace2 in the adversarial setting, is shown in Fig. 3. In the before adaptation plot, Fig. 3a the source and target domain points are close enough. This shows how similarity between the source and target domain. After adaptation, the points in the source and target domain moved farther apart, but the classification results are better after adaptation. This can be because, before adaptation, the classifier was not performing well on the classes, but after adaptation, though the source and target domain points moved apart, the classifier is able to classify both the classes irrespective of the class imbalance in the target domain data. Also, there is a significant improvement from the baseline models. But when compared to the supervised methods, their numbers are lagging a lot. Though there is room for improvement for unsupervised models compared to the supervised models, when the labels are unavailable, this is a promising direction.

## 6 CONCLUSION

In this work, we addressed an important question, can unsupervised deep domain adaptation methods be used to infer binary engagement of students in a classroom. For this, we used a discrepancy-based method and an adversarial method with different base models that work with images and videos for the experiments. The JAN model with RGB-I3D as the base network resulted in 68% accuracy and an f1-score of 0.80 performed better among JAN models. Further, using the adversarial method using Wasserstein distance resulted in the best result with 71% accuracy and f1-score of 0.82 among all the unsupervised methods. The experiments show that unsupervised domain adaptation is a promising direction for inferring student engagement when labels are not available. Moreover, the experiments showed that we can safely transfer features from an online setting to a classroom setting with unsupervised domain adaptation.

## ACKNOWLEDGEMENTS

The first author would like to thank Visvesvaraya PhD Scheme, Ministry of Electronics and Information Technology (MeitY), Government of India for supporting the work.

## REFERENCES

- Abedi, A. and Khan, S. (2021a). Affect-driven engagement measurement from videos. *arXiv preprint arXiv:2106.10882*.
- Abedi, A. and Khan, S. S. (2021b). Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network. *The 18th Conference on Robots and Vision*.
- Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17*, page 214–223. JMLR.org.
- Ashwin, T. and Guddeti, R. M. R. (2019). Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. *Education and Information Technologies*, pages 1–29.
- Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE.
- Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 67–74.
- Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.
- Drossos, K., Magron, P., and Virtanen, T. (2019). Unsupervised adversarial domain adaptation based on the wasserstein distance for acoustic scene classification. *arXiv preprint arXiv:1904.10678*.
- Geng, L., Xu, M., Wei, Z., and Zhou, X. (2019). Learning deep spatiotemporal feature for engagement recognition of online courses. In *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 442–447. IEEE.
- Grafsgaard, J., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., and Lester, J. (2013). Automatically recognizing facial expression: Predicting engagement and frustration. In *Educational Data Mining 2013*.
- Gupta, A., D'Cunha, A., Awasthi, K., and Balasubramanian, V. (2016). Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885*.
- Gupta, S. K., Ashwin, T., and Guddeti, R. M. R. (2019). Students' affective content analysis in smart classroom environment using deep learning techniques. *Multimedia Tools and Applications*, pages 1–28.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Huang, T., Mei, Y., Zhang, H., Liu, S., and Yang, H. (2019). Fine-grained engagement recognition in online learning environment. In *2019 IEEE 9th international conference on electronics information and emergency communication (ICEIEC)*, pages 338–341. IEEE.
- Liao, J., Liang, Y., and Pan, J. (2021). Deep facial spatiotemporal network for engagement prediction in online learning. *Applied Intelligence*, pages 1–13.
- Long, M., Zhu, H., Wang, J., and Jordan, M. I. (2017). Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2208–2217. JMLR. org.
- Nezami, O. M., Dras, M., Hamey, L., Richards, D., Wan, S., and Paris, C. (2019). Automatic recognition of student engagement using deep learning and facial expression. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 273–289. Springer.
- Raca, M. (2015). Camera-based estimation of student's attention in class. Technical report, EPFL.
- Thomas, C. and Jayagopi, D. B. (2017). Predicting student engagement in classrooms using facial behavioral cues. In *Proceedings of the 1st ACM SIGCHI international workshop on multimodal interaction for education*, pages 33–40.
- Thomas, C., Nair, N., and Jayagopi, D. B. (2018). Predicting engagement intensity in the wild using temporal convolutional network. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 604–610. ACM.
- Wang, M. and Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153.
- Wang, Y., Kotha, A., Hong, P.-h., and Qiu, M. (2020). Automated student engagement monitoring and evaluation during learning in the wild. In *2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, pages 270–275. IEEE.
- Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., and Movellan, J. R. (2014). The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98.
- Wilson, G. and Cook, D. J. (2020). A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46.
- Yang, J., Wang, K., Peng, X., and Qiao, Y. (2018). Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 594–598. ACM.
- Zhang, H., Xiao, X., Huang, T., Liu, S., Xia, Y., and Li, J. (2019). An novel end-to-end network for automatic student engagement recognition. In *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, pages 342–345. IEEE.