Connecting the Dots: Analysis Psychological Distress and Conversational Engagements with Transfer Learning

Elim Yi Lam Kwan (ylk25)

***Abstract*—Psychology studies indicated that distress have important implications on people's engagement in conversation [1]. However, in today's machine learning literature, most of the research focused on using low-level body language for automatic detection of signs of distress [2]. Yet, one challenge in developing a model for detecting abstract affective status (such as conversational engagement) is that it requires a lot of qualitative data.**

Therefore, we have proposed a framework for investigating conversational engagement and distress as a novel application of supervised domain adaptation. The framework transfers the knowledge learnt from students engagements in the e-learning environment (source domain) to participants engagements in well-being interviews (target domain). We have developed our e-learning engagement detector and generated an annotation guide to obtained engagement labels for the target domain. Our proposed e-learning engagement detector has successfully attained a high accuracy and F1 scores of 93.57% and 0.93, which is 7.14% and 6.45% higher than the baseline [3]. Moreover, using only 188 input tensors for training a model of 0.69M parameters, our transfer learning module achieved a reasonable accuracy and F1 of 62.50% and 0.64. Utilising the conversational engagement detector developed, we showed that engagement appeared to be correlated to an individual's distress status. Finally, we hoped that our findings would assist the development of higher performance distress classifier.

I. Introduction

With the advancement in camera technology and machine learning algorithms, using sensors data to aid the analysis of complex human emotions is not unheard of. Psychological distress is a growing issue in modern-day society and can be diagnosed by psychiatrists. Early detection of distress is often a key factor in treatment [4], moreover, the cost of ongoing assessment with human experts is prohibitive. Hence, vision-based machine learning models can potentially serve as a "second opinion". A considerable amount of literature has been published on automatic detection of signs of psychological distress based on body languages, facial expression [5] and multi-modal model [2], and are mainly based on supervised learning. However, higher-level features are also highly relevant. Psychology studies show that depression and anxiety have various impact on people's behaviour in conversation [1]. Moreover, labels related to high-level behaviour in affective datasets are often not readily available and time-consuming to collect. Therefore, we propose to learn from our machine learning model - leveraging a transfer learning approach to explore the correlation between these high-level behavioural features and psychological distress. Our framework transfers the knowledge learnt from students engagements in the e-learning environment (source domain) to participants engagements in well-being interviews (target domain). Also, our source code is available on GitHub[1].

To the best of our knowledge, this is the first paper to draw links between conversational engagements with distress through transfer learning. Our main contributions are:

- Designed a high-performance vision-based engagement detection model for e-learning environment, which outperformed the original publication by 7.14% and 6.45% in terms of accuracy and F1 scores;
- Developed a detector for abstract affective status (conversational engagement) with data efficient training based on supervised domain adaptation techniques;
- An analysis on the correlations between distress and conversational engagements.

We hope that our work can elicit interests in enhancing our understandings of psychology topics through machine learning techniques. The findings may also inspire future research to consider more abstract features for distress classification, on top of the low-level body language features widely discussed in current literature.

II. Related Work

1) Distress Detector: Firstly, we would like to provide a brief review of machine learning models for distress detection. In 2018, P. Nair and S. V [5] demonstrated their model on distress prediction based on facial expression with Naïve Bayes classifier. Moreover, they had adopted a more relaxed definition of distress, which assumed negative emotions are equivalent to distress; whereas, in our project, distress labels were supported by established psychological evaluation questionnaires. Another example of a higher complexity model for distress classification is from Lin et al. [2]. Their group established their model based on full-body data and presented a multi-modal distress classification pipeline. However, most of the published work focused on low-level action units, hence, we would like to complete these classifiers by introducing the potential of higher-level analysis data.

2) Engagement Detector: Secondly, we also evaluated the literature on the engagement detection model based on visual cues. One of the main challenges in the domain is data sparsity. With limited specialised engagement data, Nezami et al. [6]

[1]https://github.com/elimkwan/Engagements-Detector

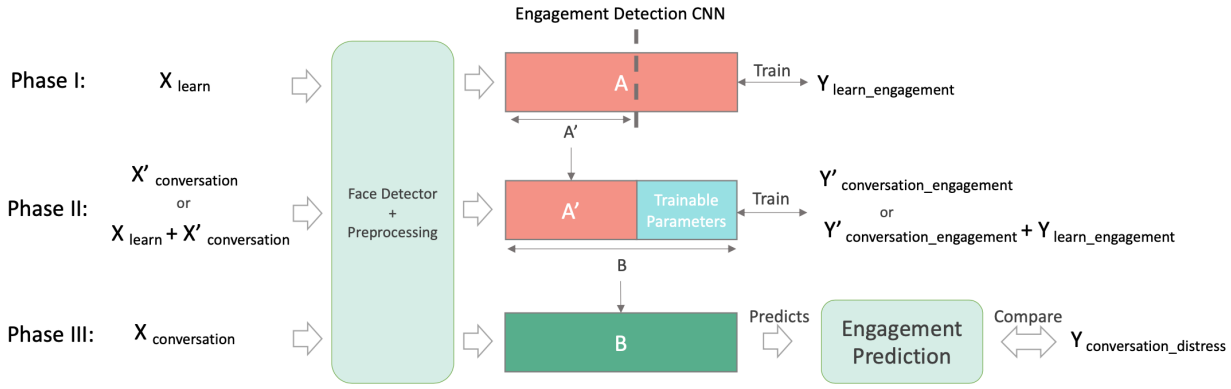


Fig. 1: Overview of the Framework

proposed to split the model by adopting two training phases: pre-trained a Convolutional Neural Network (CNN) on readily available basic facial expression data (FER-2013); then utilised those weights to initialize their engagement recognition model. Although the results seem to be promising, their model complexity is much higher than ours, with four times more parameters to be trained. Therefore, other model architectures were considered. Credited to the newly available student engagement dataset DAiSEE [7], Murshed et al. [3] had engineered a high-performance model for student engagement detection in an e-learning environment, which combines benefit from All-Convolutional-Network [8], Network-in-Network [9], and Very-Deep-Convolutional-Network [10]. In this project, we have adopted their engagement model as the baseline and improved it to better suit our application. Nevertheless, due to the contextual difference of engagement in online lessons and conversations, transfer learning techniques were applied to ensure knowledge transferred were reasonable.

3) Transfer Learning: Thirdly, transfer learning refers to machine learning techniques where a model developed for a task is reused as the starting point for a model on another task. To accomplished cross-domain knowledge transfer, domain adaptation techniques was considered to ensure model performance. In 2021, Palicki et al. (2021) [11] attempted to use transfer learning as a tool for gaining knowledge about one's affective state, they have self-annotated some data in the target domain for supervised domain adaptation. Our project framework is similar to theirs, except that we adopted more simple domain adaption strategies due to time constraints. Moreover, their emphasis on minimising domain variance also provided us some insights into our data preparation procedures.

4) Definitions and Data Annotation : Last but not least, we would like to look into various definitions and annotation schemes for distress and engagements. **Distress:** For a more relaxed definition, distress usually refers to the experiences of negative emotions, e.g. anger, sadness and fear [5]. While in medical assessments, distress can be diagnoses with PHQ-8 (depression) and GAD-7 (anxiety) indexes [12]. Participants would answer a set of multiple-choice questions and their responses would be quantified. Higher scores indicate more severe symptoms. The well-being dataset used in this project contains these indexes and classifications. **Engagements:** Most definitions describe engagement as attentional and emotional involvement in a task [13, 14]. Using the DAiSEE dataset, Murshed et al. [3] have categorised engagements in e-learning as:

- Level 0: Looking away, closed eyes
- Level 1: Eyes barely open, brow lowering
- Level 2: Mouth dimpling, eyelid tightening

Although most of these definitions could be translated to conversational engagement, e.g. direction of gaze [13], some fundamental differences remain. Interpretations on grinning, open-mouthed, head tilting in the context of learning and conversation is quite the opposite. For instance, displaying head movements and other social acts in conversations are considered as engaged [14], which is not the case in a classroom setting. In Section III, we would further discuss how we closed the gap between the two domains.

III. Methodology - Overview and Data

A. Overview

The project was divided into 3 phases as shown in Figure 1: I) Developing the engagement detection model based on students e-learning dataset, II) Adapting the model for engagement detection in conversations, III) Applying the model on the well-being dataset and analysing their level of engagement against their distress status. The engagement detector uses facial expressions of participants from 5 consecutive frames for binary classification.

B. Source Dataset

For our source domain, the DAiSEE dataset [7] was utilised for the development of the engagement detector. The dataset was intended for evaluating student engagement level in online courses. It comprises 9068 video snippets captured from 112 users - 23 females and 80 males. Each video is approximately 10 seconds and the data were collected in an unconstrained environment. It was annotated with crowdsourced labels for engagement, frustration, confusion, and boredom, and each

affective state has an associated intensity of 0 - 3 (with 3 being the highest). Since the dataset was heavily skewed and to reduce the model complexity, we have regrouped the labels. Data that was initially categorised as engagement = 3 will be considered as engaged, whereas, data that belongs to engagement level 0 or has exhibited strong negative emotions (frustration, confusion, or boredom = 3) will be considered as disengaged. They were referred to as $X_{learn}, Y_{learn\_engagement}$ in Figure 1. Table I summarised the characteristic of the two groups.

TABLE I: Exemplary Characteristic of Students' Facial Expression with respect to Engagements

Engaged	Mouth dimpling, eyelid tightening
Disengaged	Looking away, closed eyes, or strong intensity in frustration, confusion, or boredom

C. Target Dataset

For our target domain, the well-being dataset from Lin et al. [2] was adopted. It consists of full-body video interview clips collected from 35 participants. It was labelled with participant responses to self-evaluation questionnaires, which allowed us to quantify their depression and anxiety levels with well-established psychology index PHQ-8, GAD-7. One is considered to be in distress if:

$$\begin{cases}(PHQ-8>9),(GAD7>5)\\(PHQ-8>5),(GAD7>9)\end{cases}$$

These labels were referred to as $X_{conversation}$, $Y_{conversation\_distress}$ in Figure 1.

Moreover, as one of the attempts to tackle the fundamental differences of engagements in the 2 domains, we isolated the instances where the interviewees were listening to the interviewers, because talking is considered as disengaged in the classroom. In particular, we focused on the moments where the three leading questions were asked:

- Scenario 1: Describe the happiest moments in your life.
- Scenario 2: Describe the saddest moments in your life.
- Scenario 3: Describe a hobby that you used to enjoy when young but are no longer doing it now.

This approach was inspired by the transfer learning literature [11], which emphasised the importance of domain relevance and that minimising the difference in feature collection will yield performance gain.

D. Data Preprocessing

(a) Original (b) Gray-scaled (c) Hist. Equal. (d) Remove Mean

Fig. 2: Preprocessing Pipeline

In terms of data preprocessing, Figure 2 illustrated all the essential steps. The coloured frame was first transformed into a gray-scale image and the face was extracted with Viola Jones Face Detector [15]. The algorithm detected faces by extracting face-like features (Haar-like features) from the image. This involves calculating the sum of dark/light rectangular regions, which could be accomplished in real-time by utilising integral images. Also, since some of the data were collected in various illumination ranges, the faces were further processed with adaptive histogram equalization - Contrast Limited Adaptive Histogram Equalization (CLAHE) [16]. It redistributed the lightness values in the images and helped improve local contrast, which helps model learning. Moreover, CLAHE also alleviated the noise amplification challenge encountered with other histogram equalisation techniques. Its effect is shown in Figure 2c. Finally, we applied a mean removal operation as it may assist the feature extraction model [3]. Also, images were resized to 64×64 as a compromise between accuracy and model complexity. For retaining the temporal information from the videos, 3D convolution layers and sliding windows techniques will be investigated in Section IV.

E. Data Annotation for Transfer Learning

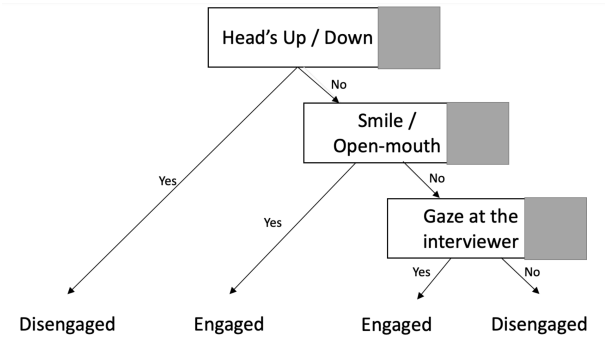


Fig. 3: Annotation Guide for Conversational Engagement[2]

Since our target domain did not come with engagement labels, we annotated a small subset of the frames from the well-being interviews. Moreover, to ensure the transfer learning pipeline is functioning properly, there are several prerequisites to be fulfilled. First and foremost, the data must be related. We argue that our domains are similar - both analysing the attentiveness of individuals using facial expression, but fundamental differences exist as previously mentioned. Therefore, we have developed our own annotation guide as shown in Figure 3, which is slightly different from the one for student engagement in Table I, and is based on literature review on gaze directions[13], head movements[14] and expression [17] in dialogue.

Frames were extracted from the first 20 seconds of the interviews when the participants had just sat down and were unlikely to be speaking. Also, none of the leading questions was asked during that duration, hence it would not overlap with the instances for the final evaluation set. In most cases,

[2]This dataset is confidential and images are included for the examinations' purposes only (lecturers and examiners)

TABLE II: E-learning Engagement CNN Model Architecture

Scheme 1	Scheme 2	Scheme 3
Five 64 x 64 Gray-scale Images		
Sliding Window	**5x1x1 Conv. 1x1 ReLU**	Sliding Window
3x3 Conv. 192 BatchNorm ReLU 3x3 Conv. 192 ReLU 3x3 Conv. 192 BatchNorm ReLU with stride 2 Dropout (0.5)	3x3 Conv. 192 BatchNorm ReLU 3x3 Conv. 192 ReLU 3x3 Conv. 192 BatchNorm ReLU with stride 2 Dropout (0.5)	3x3 Conv. 192 BatchNorm ReLU 3x3 Conv. 192 ReLU 3x3 Conv. 192 BatchNorm ReLU with stride 2 Dropout (0.5)
3x3 Conv. 192 BatchNorm ReLU 1x1 Conv. 96 ReLU 1x1 Conv. 96 ReLU	3x3 Conv. 192 BatchNorm ReLU 1x1 Conv. 96 ReLU 1x1 Conv. 96 ReLU	3x3 Conv. 192 BatchNorm ReLU 1x1 Conv. 96 ReLU 1x1 Conv. 96 ReLU
3x3 MaxPooling with Stride 2 Dropout (0.8)	3x3 MaxPooling with Stride 2 Dropout (0.8)	**3x3 Conv. with stride 2** **Dropout (0.3)**
3x3 Conv. 32 BatchNorm ReLU 3x3 Conv. 32 BatchNorm ReLU 3x3 Conv. 32 BatchNorm ReLU	3x3 Conv. 32 BatchNorm ReLU 3x3 Conv. 32 BatchNorm ReLU 3x3 Conv. 32 BatchNorm ReLU	**3x3 Conv. 32 BatchNorm ReLU** **3x3 Conv. 32 BatchNorm ReLU**
1x1 Conv. 2 global average pooling 2-way softmax	1x1 Conv. 2 global average pooling 2-way softmax	1x1 Conv. 2 global average pooling 2-way softmax

6 frames were collected from each participant, 3 for engaged, 3 for disengaged based on our annotation guide.

It is also worth noting that due to the differences in the position of the camera during the data collection process of the two datasets, front-facing students are considered as attentive, whereas participants that gaze in the direction of middle-right (where the interviewer is at) is considered as engaging, i.e the females in Figure 3 were considered as attentive.

IV. Methodology - Models

A. Engagement Models

For the engagement detector, 3 CNN architectures shown in Table II were examined.

Scheme 1 (Baseline): The baseline is based on the model proposed in Murshed et al. paper [3]. We have modified the last soft-max layer in the 3-way classifier to a 2-way softmax to fit our binary labels. The CNN was designed to provide a prediction based on a single frame. Moreover, the final decision of classification is made by accumulating the predictions over a sliding window W, and W was set as 5.

Scheme 2: The baseline utilised a sliding window to combat random noises in video data. However, we are keen to investigate whether replacing the sliding window with a 3D convolution layer would bring benefits - this forms our second scheme. It takes in an input tensor of $5 \times 64 \times 64$.

Scheme 3 (Proposed): In later experiments at Section V-B2, we discovered that the 3D convolution layer did not provide the performance gained expected, hence, we also investigated other optimisation strategies:

- Revert to the use of the sliding window.
- Replace Max-Pooling with Convolutional Layers of increased strides. It was inspired by All-CNN [8]. Their empirical studies show that a network consisted solely of convolutional layers, with occasional dimensionality reduction by using a stride could achieve state-of-the-art performance.
- Remove extra 3×3 convolutional layers at the end of the encoder to reduce the network depth. The rationale behind is that binary classification could be performed with less sophisticated features.

Additionally, the proposed model has inherited several attractive features from the Baseline, namely:

- The choice of heterogeneous blocks over homogeneous blocks which allows for a sparser network depth and thus improving computation efficiency [3].
- The use of 1×1 convolutional layers to increase non-linearity of the decision function while not impacting the receptive fields [10].
- Append batch normalisation to 3×3 convolutional layers [18].
- Replace fully connected layers with Global Average Pooling layers (GAP).

In the GAP layer, n number of feature maps were generated, where n equals to the number of classes. Afterwards, we averaged these maps and performed a softmax operation. Studies [8, 9] show that this approach reduced the number of parameters and is more transparent compared with fully connected layers, hence, improved accuracy.

These summarised our proposed scheme.

B. Transfer Learning Model

For the transfer learning model, we have inherited the encoder part of the engagement detector and re-trained the classifier, i.e. the last row in Table II. The background principle is the feature extraction part for engagement detection in the context of learning verse conversation should be similar. Moreover, their differences lie in their definitions of which feature corresponds to engaged and disengaged. Hence, to adapt our model for the target domain, the classifier part should be retrained. Our qualitative analysis in Section III and literature review in Section II-4 supported these assumptions.

In addition, if we simply apply the existing source model on our target domain, i.e. unsupervised domain adaptation, predictions' accuracy could not be quantified, which would lead to less robust findings. Therefore, we have verified 2 domain adaptation approaches: Supervised [11] and Semi-Supervised [19]. The supervised approach will only use the

target domain data for training, whereas the semi-supervised approach will use a combination of target and source domain data $(X_{conversation}, Y_{conversation\_engagement})$ and $(X_{learn}, Y_{learn\_engagement})$ for training. Both models will be evaluated on the target domain data $X_{conversation}, Y_{conversation\_engagement}$ only. The impact of semi-supervised training is two-folded. On one hand, there are only limited self-annotated data for the target domain (188 sets of five images in total), which may be insufficient for a supervised domain adaptation approach. On the other hand, the dissimilarities between the definitions of engagement in the source and target domain will cause the model to learn the incorrect knowledge. Their respective performance will be thoroughly evaluated in the Section V-C. Figure 1 also provided an illustrated of the above concepts.

V. Experiments and Results

Our proposed e-learning engagement detector has attained a high accuracy and F1 scores of 93.57% and 0.93, which is 7.14% and 6.45% higher than the original publication [3]. Furthermore, the transferred model has also attained 62.50% and 0.64 accuracy and F1 scores in the target domain. Moreover, applying the final conversation engagement detector on the well-being interview data shows that engagement seems to be correlated with distress, which coincided with our hypothesis.

A. Performance Matrices

3 matrices will be used to evaluate the performance of the models: Accuracy, F1 Scores and Confusion Matrix. Credit to the regrouping discussed in Section III-B, we could obtain a balanced dataset. Moreover, the F1 Score was still adopted as it allows us to summarise a classifier general performance in terms of Precision (P) and Recall (R). It was being computed as follows:

$$P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN} \tag{1}$$

$$F1 = 1 \times \frac{P \times R}{P + R} \tag{2}$$

where TP, TN, FP, FN refers to True Positive, True Negative, False Positive, False Negative.

B. Engagement Model

1) Model Training: The proportion of (training + validation): (testing) dataset was 7:1. For the data splits, we retained the original splits of the DAiSEE dataset, where they have ensured all splits are mutually exclusive and exhaustive with respect to subjects. Additionally, we applied k-fold cross-validation with $k = 5$. Each round, 358 sets of five images were applied for training, while 90 sets of five images were used for validation. Adam was selected as the optimiser due to its adaptive step size, and loss is based on binary cross-entropy as this is a 2-class classification task. Mini-Batching was used to improve the speed of the network convergence. The model was trained with the learning rate 0.001, batch size 32 and 50 epochs (with early stop and patience of 10). Moreover, the loss

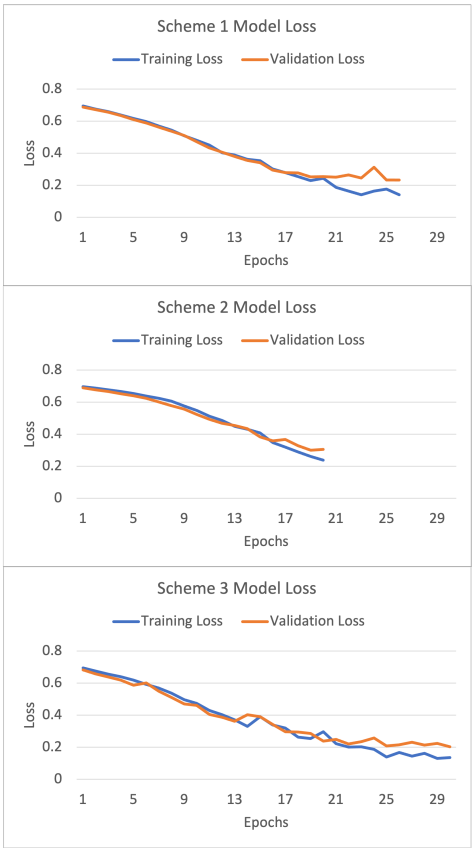


Fig. 4: Model Train and Validation Loss

of training and validation data over epochs were monitored closely, we early stopped once the validation loss is higher than the minimum validation loss for 10 consecutive frames. This helps avoid over-fitting and improves training efficiency. Finally, to compare performance among models, we choose the models with the highest validation accuracy among the folds and measure their performance with the test set. This ensures there is no data leakage.

Figure 4 shows the training details and the convergence rate of the 3 schemes. Scheme 3 took the longest time to converge, whereas Scheme 2 is the quickest. This could be understood as Scheme 3 has 12.22% more parameters than Scheme 2 according to Table III. Their accuracy, reliability and coverage will be further discussed in Section V-B2.

TABLE III: Engagement Detector Accuracy and F1 Scores

Schemes	Accuracy (%)	F1	# Parameters
1 (Baseline)	87.50	0.8824	602,946
2	75.00	0.7500	602,952
3 (Proposed)	93.75	0.9393	676,674

TABLE IV: Confusion Matrix for Scheme 1

Engagement Levels	0	1
0	40.63%	9.38%
1	3.13%	46.88%

TABLE V: Confusion Matrix for Scheme 2

Engagement Levels	0	1
0	37.50%	12.50%
1	12.50%	37.50%

TABLE VI: Confusion Matrix for Scheme 3

Engagement Levels	0	1
0	45.31%	14.06%
1	1.56%	48.44%

2) Model Performance : Our proposed e-learning Engagement Detector has attained a high accuracy and F1 scores of 93.57% and 0.93, which is 7.14% and 6.45% higher than the baseline [3]. The performance of the models was summarised in Table III. This could be attributed to the replacement of max-pooling layers with convolutional layer and the reduced network depth. The empirical data also support our hypothesis that binary classifier requires less sophisticated features compared with multi-class classifiers. As shown in Table III, the removal of the last layer of the encoder results in a positive gain in performance. Nevertheless, there exists a trade-off between computational efficiency and performance, Scheme 3 attained the highest accuracy but uses 12% more parameters compared with Scheme 1. Moreover, F1 scores and the confusion matrices **??** reflected that all the classifiers have good recall and precision, and that true positive and true negative data always prevails.

Another observation is that using 3D convolution layers as a replacement for the sliding window operation did not bring any positive gain in our application. This may be due to the early placement of the 3D convolutional layer, as it compressed the 5 consecutive frames into 1 at the very start. Hence, if a lot of noises were to presence in any of the frames, misclassification will occur. This setup does not promote risk diversification, which is essential in combating random sensor noise. Thus, it performed the worst. Compared with the baseline, it uses 0.99% more parameters and obtained 12% lower accuracy and F1 Scores. In the future, we could consider the use of the LSTM model to fully utilised the temporal data.

C. Transfer Learning

TABLE VII: Data Distribution for Semi-Supervised Domain Adaptation

	Train	Val.	Test
Target Domain (sets of 5 images)	152	20	24
Source Domain (sets of 5 images)	100	14	-

TABLE VIII: Transferred Models Accuracy and F1 Scores

	Accuracy (%)	F1
Supervised	62.50	0.64
Semi-Supervised	45.83	0.13

2 domain adaptation approaches, supervised and semi-supervised were examined. We trained the parameters with cross-validation (k = 5), the learning rate of 0.0005 and batch size of 1. It was also trained with 50 epoch but with early stopping and patience of 5. The model that achieved the best validation accuracy among the folds will be chosen, and its performance on the test set is as shown in Table VIII. The (test) : (train + validation) split is 1:7. For the semi-supervised approach, the proportion of target to source data is 1.54 : 1, Table VII shows a more detailed breakdown.

TABLE IX: Confusion Matrix for Supervised Domain Adaptation

Engagement Levels	0	1
0	29.17%	20.83%
1	16.67%	33.33%

TABLE X: Confusion Matrix for Semi-Supervised Domain Adaptation

Engagement Levels	0	1
0	41.67%	8.33%
1	45.83%	4.17%

Our supervised domain adaptation model has attained 62.50% and 0.64 accuracy and F1 scores in the target domain. Although it is not particularly high, it is reasonable as the target domain data is quite sparse. This model will be used as the final model for the analysis in Section V-D. Future work should consider improving the predictor by:

- Increasing the amount of target domain labels;
- Improving the quality of the target domain labels with more sophisticated features(e.g. content of the conversation) and annotation schemes;
- Replacing the source domain dataset with one that is more relevant to conversational engagement.

Another interesting trend that relates to the comments made in Section IV-B is that the performance of the supervised approach is much more superior than that of the semi-supervised approach. As shown in the confusion matrix in Table X, the semi-supervised approach failed to learn and has a low recall. This can be explained as the differences in the definition of engagements in the 2 domains would confuse the model. We had attempted to resolve the challenge by having a higher proportion of source domain data, unfortunately, the differences remain.

D. Correlation between Distress and Engagement

Finally, the transferred model was applied to the well-being interview data as a tool for engagement prediction. We captured the participants' engagement levels when questions related to happiness, sadness and remembrance were asked. Some intriguing insights could be drawn.

Firstly, it is observed that individual with a high level of distress tends to have low engagements in all the scenarios. In Figure 5, 15, 14 and 13 participants who have high distress levels displayed signs of low engagements in the 3 scenarios respectively, which is the highest among all the combinations of distress-engagement levels.

Secondly, Figure 6 indicated that engaged individual tends to have lower depression and anxiety scores. Most engaged

Fig. 5: Engagements when being asked about Happiness / Sadness / Remembrance

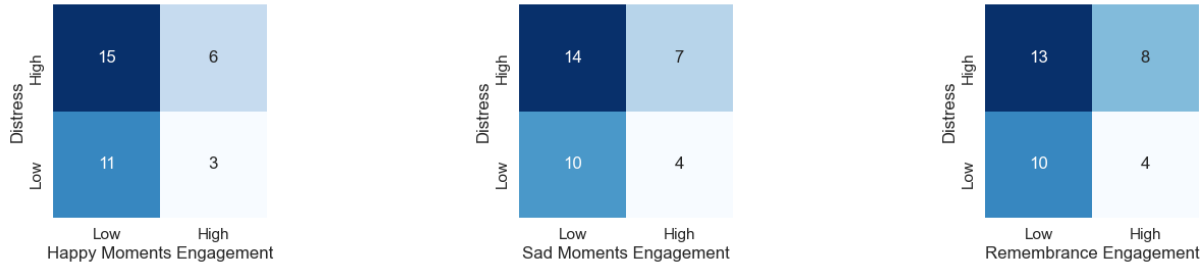
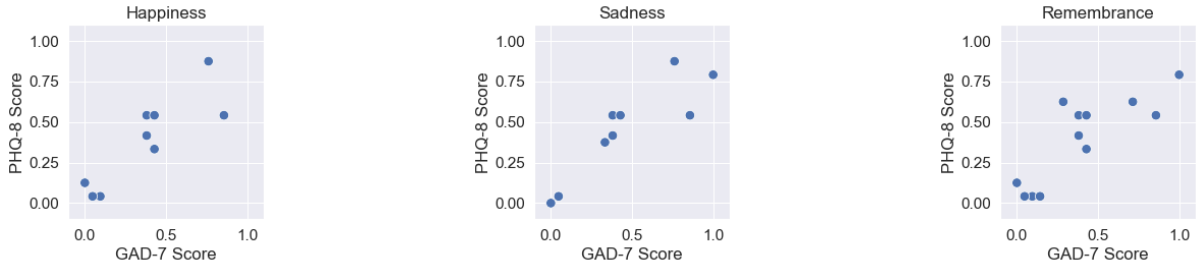


Fig. 6: Scores of Engaged Individual when being asked about Happiness / Sadness / Remembrance



participants have normalised PHQ-8 and GAD-7 below 0.75 and 0.5. This supported the trends observed in Figure 5.

However, we also noticed that the engagement predictor tends to classify one to be of low engagement. This intrinsic model bias may have impacted the robustness of the findings. Hence, future work should focus on improving the performance of the classifier.

To conclude, level of engagements appears to be correlated with distress. More thorough experiments will be required to confirm their relationship. Moreover, the framework proposed could be a starting point for future research in using transfer learning for sentiment analysis.

VI. Conclusion

We introduced a high-performance vision-based Engagement CNN Model in the context of e-learning, which has attained accuracy and F1 scores of 93.57% and 0.93, a 7.14% and 6.45% improvements compared with the baseline. Furthermore, we proposed a novel framework for drawing links between conversational engagements with distress through supervised domain adaptation. Credit to the domain adaptation techniques employed, our transfer learning module achieved a reasonable accuracy and F1 of 62.50% and 0.64 with limited self-annotated data (188 input tensors for training a model of 0.69M parameters). Finally, an analysis on high-level behaviour features (conversational engagements) was carried out, which provided intriguing insights on the correlation between engagements and distress.

Our research demonstrated the multi-folded potential of machine learning in the field of affective computing. It could also be inspirational for how understandings on psychiatry topics could be gained through transfer learning, which reaffirms findings from psychology studies and empirical medical assessments. Lastly, our results may also inspire future research to consider incorporating higher-level behaviour symptoms as one of the features in multi-modal distress classifier, in addition to the low-level body language features that have been actively researched.

VII. Limitations and Future Work

Last but not least, we would like to acknowledge some limitations of the project and proposed some future directions:

- Domain differences: The transfer learning model could be improved by adopting a more relevant source dataset. Moreover, datasets related to affective states are usually private and hard to obtain, hence, given the limited time of the project, DAiSEE was adopted despite their differences in context. Future work could consider other engagement datasets that are more relevant to conversations.
- Bias induced through self-annotation: Since all the target domain engagement labels were annotated by a single individual, bias may exist. Moreover, we had attempted to minimise such risk by crafting an objective and detailed annotation roadmap.
- More advanced temporal model could be adopted instead: Sliding window had provided adequate performance in this project. If more time were allowed, we could investigate the optimum size of the sliding window as well. Moreover, if the framework were to scaled to multi-class engagement classification, a more advanced model, such as LSTM could also be considered.

Although limitations exist, we hope that our research has opened the doors for more work to explore the use of transfer

learning in the domain of affective computing as a mean of tackling the data sparsity challenge.

VIII. Acknowledgement

The author would like to thank Dr Marwa Mahmoud for her guidance and provision of the well-being dataset. The author would also like to express gratitude to the creator of the DAiSEE dataset [7], which made it public available for research purposes.

References

[1] Leanne K. Knobloch, Lynne M. Knobloch-Fedders, and C. Emily Durbin. "Depressive Symptoms and Relational Uncertainty as Predictors of Reassurance-Seeking and Negative Feedback-Seeking in Conversation". In: *Communication Monographs* 78.4 (2011), pp. 437–462. DOI: 10.1080/03637751.2011.618137.

[2] Weizhe Lin et al. *Looking At The Body: Automatic Analysis of Body Gestures and Self-Adaptors in Psychological Distress*. 2020. arXiv: 2007.15815 [cs.CV].

[3] M. Murshed et al. "Engagement Detection in e-Learning Environments using Convolutional Neural Networks". In: *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*. 2019, pp. 80–86. DOI: 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00028.

[4] E. Jane Costello. "Early Detection and Prevention of Mental Health Problems: Developmental Epidemiology and Systems of Support". In: *Journal of Clinical Child & Adolescent Psychology* 45.6 (2016). PMID: 27858462, pp. 710–717. DOI: 10.1080/15374416.2016.1236728.

[5] P. Nair and S. V. "Facial Expression Analysis for Distress Detection". In: *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. 2018, pp. 1652–1655. DOI: 10.1109/ICECA.2018.8474761.

[6] Omid Mohamad Nezami et al. *Automatic Recognition of Student Engagement using Deep Learning and Facial Expression*. 2019. arXiv: 1808.02324 [cs.CV].

[7] Abhay Gupta et al. *DAiSEE: Towards User Engagement Recognition in the Wild*. 2018. arXiv: 1609.01885 [cs.CV].

[8] Jost Tobias Springenberg et al. *Striving for Simplicity: The All Convolutional Net*. 2015. arXiv: 1412.6806 [cs.LG].

[9] Min Lin, Qiang Chen, and Shuicheng Yan. *Network In Network*. 2014. arXiv: 1312.4400 [cs.NE].

[10] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV].

[11] Sean-Kelly Palicki et al. *Transfer Learning Approach for Detecting Psychological Distress in Brexit Tweets*. 2021. arXiv: 2102.00912 [cs.CL].

[12] Oncology Nursing Society. *PHQ and GAD7 Instruction Manual*. Tech. rep.

[13] Yukiko I. Nakano and Ryo Ishii. "Estimating User's Engagement from Eye-Gaze Behaviors in Human-Agent Conversations". In: *Proceedings of the 15th International Conference on Intelligent User Interfaces*. IUI '10. Hong Kong, China: Association for Computing Machinery, 2010, pp. 139–148. ISBN: 9781605585154. DOI: 10.1145/1719970.1719990. URL: https://doi.org/10.1145/1719970.1719990.

[14] Loredana Cerrato and Nick Campbell. "Engagement in Dialogue with Social Robots". In: *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*. Ed. by Kristiina Jokinen and Graham Wilcock. Singapore: Springer Singapore, 2017, pp. 313–319. ISBN: 978-981-10-2585-3. DOI: 10.1007/978-981-10-2585-3_25. URL: https://doi.org/10.1007/978-981-10-2585-3_25.

[15] P. Viola and M. Jones. "Rapid object detection using a boosted cascade of simple features". In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*. Vol. 1. 2001, pp. I–I. DOI: 10.1109/CVPR.2001.990517.

[16] G. Yadav, S. Maheshwari, and A. Agarwal. "Contrast limited adaptive histogram equalization based enhancement for real time video system". In: *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. 2014, pp. 2392–2397. DOI: 10.1109/ICACCI.2014.6968381.

[17] Dan Bohus and Eric Horvitz. "Models for Multiparty Engagement in Open-World Dialog." In: Jan. 2009, pp. 225–234. DOI: 10.3115/1708376.1708409.

[18] Sergey Ioffe and Christian Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 2015. arXiv: 1502.03167 [cs.LG].

[19] Ting Yao et al. "Semi-Supervised Domain Adaptation With Subspace Learning for Visual Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015.