Advancing Online Education: An Artificial Intelligence Applied System for Monitoring and Improving Employee Engagement in Enterprise Information Systems

Nguyen Thanh Son[1], Trong Tien Hoang[2], Satyam Mishra[2], Nguyen Thi Bich Thuy[3], Tran Huu Tam[4], and Cong-Doan Truong[2(✉)]

[1] VNPT-Media Software Company, Hanoi, Vietnam
sonnt.196@gmail.com
[2] International School, Vietnam National University, Hanoi, Vietnam
{tienht,tcdoan}@vnu.edu.vn
[3] University of Science, Vietnam National University, Hanoi, Vietnam
nguyenthibichthuy@hus.edu.vn
[4] FPT University, Hanoi, Vietnam
tamth3@fe.edu.vn

Abstract. Online learning has gained significant popularity, but maintaining learner focus remains a challenge, especially in financial enterprise training systems. The need for training has increased with banking and finance digitalization trends, yet high learning curves and prolonged sessions often lead to distractions. This research introduces an online learning tool that monitors and quantifies learner attention in real-time. Using the MobileNet Convolutional Neural Network, we detect seven core emotions, which, combined with attention scores, form a Concentration Index (CI). Learners are then categorized as "Highly-engaged," "Normally Engaged," or "Disengaged." With 70% accuracy on training and 65% on testing, our engagement metrics provide actionable insights for educators and administrators, enhancing virtual learning and aiding in analytical problem-solving strategies.

Keywords: Online Learning · Learner Focus · Real-Time Attention Quantification · Convolutional Neural Network (CNN) · MobileNet · Emotion Detection · Concentration Index (CI)

1 Introduction

Enterprises are continuously evolving, adopting advanced technologies to improve customer experience. As a result, there's an imperative to ensure the workforce is updated with the latest knowledge to adeptly handle these innovations. Many financial institutions have adopted in-house Learning Management Systems (LMS) to foster knowledge transfer, enhance student engagement, and utilize results for career progression. Current

P. Cong Vinh and H. Mahfooz Ul Haque (Eds.): ICTCC 2023, LNICST 586, pp. 3–18, 2024.
https://doi.org/10.1007/978-3-031-59462-5_1

forecasts anticipate the global LMS market to grow at a rate of 23% by 2024, reaching an estimated revenue of roughly 12.48 billion [1]. However, these enterprise LMS systems are often standalone systems in the enterprise system architecture or only providing very high-level results on learning performance of the attendants.

E-learning holds the promise of personalizing learning experiences, delivering real-time feedback to educators on student performance [2]. Its initial adoption at the university level showcased its merits, where traditional full-time onsite learning was often perceived as time-consuming, whereas online methods were lauded for their flexibility and cost-efficiency [3]. Yet, e-learning poses challenges in areas like peer and educator-student interactions, prompt educator support, and sustaining attention [4]. As its popularity soars, these issues become more evident, driving our aim to innovate in emotion detection and ongoing student engagement assessment.

Various studies underscore the individual's ability to grasp cognitive processes [5, 6]. Some emphasize the creation of learning systems that discern learners' emotional states [7–9]. Other research explores integrating motivational skills in smart learning systems [10]. With evidence supporting emotion detection via facial expressions, it's been found that six unique emotions can predictably be identified with 89% accuracy [11]. While one study [12] presented in-depth results on facial expression-based emotion recognition, understanding emotions like boredom and fatigue remains under-explored. A segment of research champions the identification of confidence, frustration, and boredom in e-learning to deliver tailored feedback [13]. The value of embedding emotions in e-learning was asserted in [14]. Notwithstanding extensive studies on eye movement [15, 16] and yawn detection [17], tracking drowsiness and unusual head movement has been less addressed in e-learning contexts. Though past work has touched upon emotion recognition in e-learning, a comprehensive solution with a proof of concept has been absent.

Our study endeavors to create a comprehensive online learning platform incorporating emotion recognition. This solution encompasses an Android app fostering educator-student engagement, a desktop web interface, and robust computing units and databases for machine learning, feature storage, and data management. The aim is not only to improve in-class learning performance, but also to provide management with novel quantifiable metrics in assessing employee performance such as ability to concentrate over a long period of time, ability to learn new skills.

- Step 1: Face Recognition - The product starts by integrating the MTCNN (Multi-Task Cascaded Convolutional Neural Network) model for face recognition. This foundational step paves the way for tracking learners' facial expressions in subsequent stages.
- Step 2: Focus Classification - After successful face detection, the system utilizes the MobileNet model to determine if learners are focused. MobileNet, known for its efficient image classification prowess, is employed for this purpose.
- Step 3: Emotion Classification - The system then uses MobileNet to classify learners' emotions based on facial expressions, identifying states like happiness, sadness, surprise, among others.
- Step 4: Concentration Index Calculation - As elaborated in Sect. 2.4 of this research, a concentration index is computed to quantify engagement levels. It integrates results

from both emotion and focus evaluations to provide a holistic measure of learners' attention during e-learning.

– Step 5: Results Presentation for Educators - Lastly, the system offers a visual summary of learners' emotions, focus, and concentration indices to teachers or administrators. These real-time insights empower educators to adapt their instruction based on learners' engagement.

The product's efficiency depends on its precision in emotion recognition, focus classification, and concentration index computation. Preliminary results show a 70% success rate on the training set and 65% on the test set in detecting emotions, underscoring its potential for practical use.

2 Methodology

To accomplish the specified objective, the authors utilized the FER2013 dataset in tandem with the Multi-Task Cascaded Convolutional Neural Network (MTCNN) and MobileNet. These models' outputs were then integrated to compute the Concentration Index.

2.1 FER2013 Dataset

For retraining the MobileNet model, the FER2013 dataset [18] was selected. This dataset is apt for emotion recognition tasks executed via algorithms akin to MobileNet. It consists of grayscale facial images with pre-assigned labels. For this study, the dataset was partitioned into 28,709 training observations and 3,589 test observations (Fig. 1).

Fig. 1. A 5 × 5 collage showcasing diverse facial expressions from the FER-2013 dataset.

To address issues of limited sample size and class imbalance, the authors applied data augmentation and transfer learning techniques. Theoretically, leveraging these techniques on this dataset should bolster the overall accuracy and robustness of real-time emotion detection. Recent studies have shown that facial emotion detection tasks yield the highest accuracy when deep learning algorithms are employed. Some research even indicates enhanced accuracy with a combined dataset approach [19].

2.2 MultiTask Cascaded Convolutional Neural Network (MTCNN)

The MultiTask Cascaded Convolutional Neural Network (MTCNN) [20] is a state-of-the-art face detection technique integral to this research. MTCNN distinguishes itself by implementing a three-stage cascaded structure to optimize face detection. It commences by resizing the image across multiple scales, ensuring the detection of faces irrespective of their size.

- Proposal Network (P-network): This is the first phase, designed to roughly identify possible face regions. To achieve a comprehensive scan, the P-network uses a lower threshold which might lead to some false positives.
- Refine Network (R-network): This subsequent stage sharpens the earlier detections by refining face regions and reducing false positives. It further utilizes non-maximum suppression (NMS) to eliminate overlapping detections.
- Output Network (O-network): As the final layer, the O-network perfects the bounding box predictions, ensuring the accuracy of face detections.

One of the standout features of MTCNN is its capability to identify facial landmarks, which is invaluable for precise tasks like face alignment, further underscoring its adaptability and precision in face-related tasks. MTCNN serves as the first critical step in the pipeline of real-time emotional reaction detection. Its robustness and accuracy in face detection set the stage for subsequent models, like MobileNet, to interpret the nuanced changes in facial expressions and categorize emotional states effectively.

2.3 MobileNet

Google's MobileNet, designed with the Depthwise Separable Convolution (DSC) technique, revolutionized the field of efficient deep learning. Particularly recognized for its compactness and computational efficiency, MobileNet [21] stands out as an optimal neural network architecture for resource-limited platforms like mobile devices. The unique implementation of depth-wise separable convolutions divides the convolution process into depth-wise and point-wise layers, ensuring reduced computational overhead and fewer parameters. When applied to tasks like real-time emotion recognition using the FER-2013 dataset, MobileNet's efficiency and lightweight design make it an ideal choice, enabling fast and accurate analysis without the need for extensive computational resources.

MobileNet is specially designed to ensure efficient feature extraction while conserving computational resources. This makes it ideal for real-time applications, notably in the realm of face recognition and emotion analysis. Given the challenges of detecting

student engagement in online learning scenarios, the MobileNet architecture offers a compelling balance of performance and computational efficiency for mobile devices.

The Depthwise Separable Convolution divides the convolution process into two distinct layers [22]:

Depthwise Convolution: This applies a unique filter to each individual input channel, distinguishing it from the standard convolution where filters are incorporated across all input channels. The mathematical representation for this is:

$$\text{Depthwise } (F, K) = F \times K \times K$$

where:

F is the feature map size.

K is the kernel size.

Point-wise Convolution: Utilizes a 1×1 convolution to modify the dimensionality of the combined channels.

$$\text{Pointwise } (F, M, N) = F \times F \times M \times N$$

where:

F is the feature map size.

M is the number of input channels.

N is the number of output channels.

When analyzing the computational efficiency, the total cost for Depthwise separable convolutions is given by:

$$\text{TotalCost} = (D_F \times D_F \times M \times D_K \times D_K) + (D_F \times D_F \times M \times N)$$

where:

DF is the feature map size.

M is the number of input channels.

N is the number of output channels.

DK is the kernel size.

Compared to the computational expense of a standard convolution, the reduction achieved through Depthwise Separable Convolution can be expressed as:

$$\frac{\textit{Standard Convolution Cost} - \textit{Depthwise Separable Convolution Cost}}{\textit{Standard Convolution Cost}}$$

Within the MobileNet architecture, while there are up to 30 layers present, the core processes involve the following pivotal layers:

- Convolution Layer
- Depthwise Layer
- Pointwise Layer
- Softmax Layer: Primarily for classification.

Batch Normalization (BN) and ReLU: Upon contrasting the performance of a 30-layer network employing traditional convolution against a similar 30-layer network utilizing Depthwise Separable Convolution (MobileNet) on the ImageNet dataset, the insights reveal: MobileNet exhibits a marginal 1% drop in accuracy. However,

there's a noteworthy reduction, approximately 90%, in the Mult-Adds and parameters, underscoring its efficiency.

In the context of our task – detecting emotional reactions in the FER-2013 dataset – this efficiency means the model can quickly process and interpret facial expressions, making it a fitting choice for real-time student engagement detection.

2.4 Concentration Index Calculation

The concentration index [23] serves as a metric to gauge students' engagement during e-class learning sessions. Derived from both emotional recognition outcomes and attention scores, this real-time calculation offers insights into a student's immediate focus and attentiveness. Building the concentration index involves two key components: the emotional recognition model and the attention score. The former leverages a pre-trained convolutional neural network [24] to discern a range of emotions, such as anger, happiness, and sadness. Concurrently, eye-tracking technology maps the student's retinal focus and gaze patterns, providing data for the attention score.

By integrating emotional and attentional metrics, the concentration index offers a nuanced view of student engagement. Educators can then categorize this engagement into three levels: high, normal, and disengaged. This comprehensive approach ensures a more accurate representation of a student's involvement in the learning process.

2.5 Proposed Model Architecture

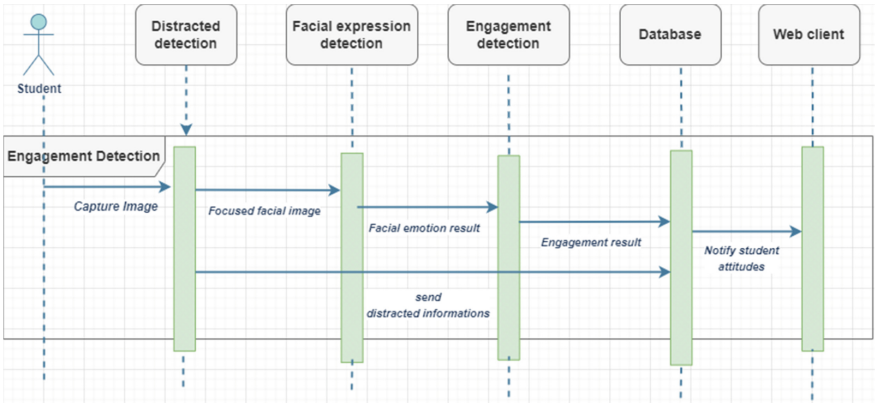

Fig. 2. Proposed Model Architecture

The proposed model architecture (Fig. 2) presents a systematic approach for real-time engagement detection in online learning settings. It adeptly merges several components—face recognition, emotion analysis, attention tracking, and the concentration index calculation—to offer a thorough insight into students' engagement dynamics.

For optimal performance and real-time responsiveness, certain tasks should be offloaded to servers with robust computational capabilities, while others should be processed at the mobile edge. Face recognition and emotion analysis, given their complexity and reliance on neural networks like MTCNN and MobileNet, would benefit from server-side processing. This would ensure high accuracy without straining the mobile device. Conversely, attention tracking, using tools like SeeSoSDK [25], and basic preprocessing tasks could be executed at the mobile edge. This division of tasks would minimize latency, particularly for operations that require immediate feedback, such as monitoring a student's gaze in real-time.

3 Results

3.1 Experiment 1: Data Augmentation Techniques

The initial experiment delves into the influence of data augmentation techniques on dataset expansion and the corresponding impact on model performance. By employing augmentation methods like horizontal flips, rotations, translations, and zoom on the training dataset [26], this experiment seeks to elucidate their role in enlarging the dataset and fortifying the model's generalization potential. Our findings shed light on how augmentation impacts the model's accuracy and its capacity to avert overfitting. While horizontal data flipping stands out as an efficient technique for data augmentation in CNNs [27], it's crucial to be mindful that for diminutive images, this method can significantly alter the image structure, potentially compromising performance (Table 1).

Table 1. ImageDataGenerator settings summary

Settings	Effect
width_shift_range = 0.15 height_shift_range = 0.15	These dictate horizontal and vertical image shifts. Considering the images are cropped to a 48x48 size centered on the face, excessive shifts might inadvertently introduce noise, potentially obscuring crucial facial features. An optimal threshold of 0.15 is established
zca_whitening	This has been disabled, given that grayscale images don't significantly benefit from this form of whitening
zoom_range = 0.15	Administers random image zooms. Any zoom constants beyond 0.15 risk distorting image proportions, possibly losing emotion-specific details
horizontal_flip = True	Activated to handle the predominance of upright faces in the dataset, this setting amplifies training diversity through mirrored image generation
shear_range = 0.15	Modifies the image by adjusting its contour angle. Maintaining a shear angle of 15 degrees ensures image proportions remain intact and complements the data augmentation strategy

In summation, the chosen augmentation settings are calibrated to balance between retaining image fidelity and emotional nuances while simultaneously enhancing training variety. Techniques like horizontal flips, rotations, translations, and zooming are tactically deployed to enrich the dataset. For example, horizontal flipping introduces mirrored image variations, mitigating any inherent orientation bias. Meanwhile, controlled manipulations through rotations, translations, and zooming bolster the model's resilience against spatial alterations. It's essential to recognize that while augmentation fosters diversification, it can also be a source of noise. Consequently, striking a balance becomes especially important and difficult. The overarching goal of Experiment 1 is to utilize these augmentation strategies to elevate MobileNet's efficacy in real-time engagement detection within online learning environments.

3.2 Experiment 2: Class Weights for Balancing Classes

Class weighting is a technique used to address class imbalances by attributing distinct weights to each class according to the number of training samples present. As delineated in the Sklearn documentation, the class weights for every class can be calculated using the formula:

$$CW = \frac{n\_samples}{n\_classes.np.bincount(y)}$$

where:

n_samples: the total number of rows in a dataset.

n_classes: the total distinct classes within the dataset.

np.bincount(y): the count of each specific class in the dataset.

The second experiment focuses on understanding the influence of class weighting on the precision of engagement detection. It encompasses the assignment of greater weights to less represented classes in line with their sample proportions. The objective is to amplify the model's responsiveness to minority classes, thus highlighting enhanced efficacy, particularly when identifying rarer emotions or engagement tiers. In challenging the predisposition towards predominant classes, the research aims for a more precise and holistic evaluation of engagement, shedding light on student dynamics during online education.

3.3 Experiment 3: Fine-Tuning MobileNet

The third experiment delves into fine-tuning the MobileNet architecture, emphasizing transfer learning for emotion recognition within online learning contexts. Utilizing the MobileNet structure pre-trained on the ImageNet dataset, a specific approach was employed for fine-tuning. In convolutional neural networks (CNNs), the latter layers discern high-level features, making them prime candidates for transfer learning [28]. Given the diminutive image size, utilizing the full extent of the MobileNet can be excessive; hence, the architecture is truncated at the 12th block. The terminal layer of the pre-trained model was discarded, paving the way for the integration of a global max-pooling layer followed by a dense prediction segment within the CNN:

– The GlobalAveragePooling2D layer amasses the distinct features of each image into a consolidated 1280-element vector.
– A dense layer is incorporated to channel these features into individual predictions for each image. A softmax activation function is paired with this dense layer, producing seven outputs to correspond with the seven identifiable emotions.
– Optimization was carried out using the categorical cross-entropy loss function paired with the Adam optimizer.

For fine-tuning, the foundational layers of the neural network were set immutable, allowing exclusive training of the terminal layer, yielding a more accurate model rendition. The applied callback functions included:

– EarlyStopping: Ceases model training if a stipulated improvement threshold isn't met over a defined epoch span.
– ReduceLROnPlateau: Curtails the learning rate when the model ceases to learn.

Training of this adapted model was conducted over the FER-2013 dataset across five epochs, as depicted in Fig. 3.

```
Epoch 1/5
449/448 [==============================] - 22s 49ms/step - loss: 0.3028 - accuracy: 0.3423 - val
_loss: 1.8122 - val_accuracy: 0.3761
Epoch 2/5
449/448 [==============================] - 22s 49ms/step - loss: 0.2968 - accuracy: 0.3656 - val
_loss: 2.1448 - val_accuracy: 0.3553
Epoch 3/5
449/448 [==============================] - 22s 50ms/step - loss: 0.2970 - accuracy: 0.3679 - val
_loss: 1.9577 - val_accuracy: 0.3892
Epoch 4/5
449/448 [==============================] - 22s 50ms/step - loss: 0.2983 - accuracy: 0.3685 - val
_loss: 2.1780 - val_accuracy: 0.3608
Epoch 5/5
449/448 [==============================] - 22s 50ms/step - loss: 0.2979 - accuracy: 0.3671 - val
_loss: 1.8160 - val_accuracy: 0.3789
```

Fig. 3. Model summation's simple fine-tuning after five epochs

Fine-tuning of the entire custom model with a learning rate: 0.0001 (Fig. 4).

```
Number of layers in the base model:  81
Model: "functional_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_2 (InputLayer)         [(None, 48, 48, 3)]       0
_________________________________________________________________
mobilenet_trunc (Functional) (None, 1, 1, 1024)        2162880
_________________________________________________________________
global_average_pooling2d (Gl (None, 1024)              0
_________________________________________________________________
pred (Dense)                 (None, 7)                 7175
=================================================================
Total params: 2,170,055
Trainable params: 2,152,263
Non-trainable params: 17,792
_________________________________________________________________
```

Fig. 4. Module summarization for secondary experiments

With 44 epochs, the authors get accuracy for seven emotion categories (Fig. 5).

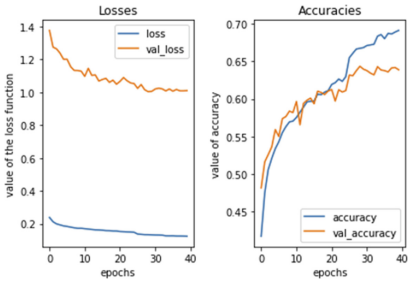

Fig. 5. Training history

3.4 Experiment 4: Analyzing Engagement Detection

The concluding experiment emphasizes the evaluation of engagement detection performance. To assess the model's precision in emotion recognition, performance metrics are calculated across each emotion category. Various visual representations, such as confusion matrices, emotion prediction visualizations, and juxtapositions of predicted versus actual labels, encapsulate the results. These graphical insights spotlight the model's proficiency and pinpoint potential areas for refinement concerning engagement level detection. Figure 6 delineates the performance metrics for the classification model. The performance is categorized for seven emotions, as follows: 0: 'Angry', 1: 'Disgust', 2: 'Fear', 3: 'Happy', 4: 'Sad', 5: 'Surprise', 6: 'Neutral'.

	precision	recall	f1-score	support
0	0.54	0.59	0.56	467
1	1.00	0.02	0.04	56
2	0.54	0.39	0.45	496
3	0.83	0.87	0.85	895
4	0.55	0.50	0.53	653
5	0.79	0.72	0.75	415
6	0.53	0.70	0.60	607
accuracy			0.64	3589
macro avg	0.68	0.54	0.54	3589
weighted avg	0.65	0.64	0.63	3589

Fig. 6. Measuring the performance of a classification model

Comparatively, our custom model showcased positive accuracy when compared against earlier models trained on the FER2013 dataset, as illustrated in Fig. 6. With a compact footprint of approximately 25 MB, the trained model is highly suitable for mobile application integrations (Fig. 7 and Table 2).

Analyzing Incorrect Predictions: The confusion matrix accentuates the model's exceptional prowess in discerning happiness, while also highlighting indicators for enhancement in other emotion categories. A confluence of factors, such as a limited dataset for certain categories and potential inconsistencies in data quality, play a part in the overall accuracy. More often than not, misclassified images bear a stronger resemblance to the predicted class than to their designated categories.

Table 2. Mislabeled cases breakdown

Category	Missed labels	Observations
Angry	238	467
Disgust	0	56
Fear	164	496
Happy	163	895
Sad	269	653
Surprise	78	415

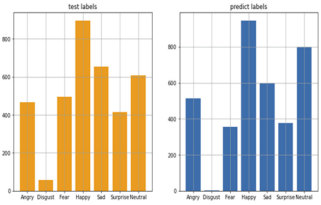

Fig. 7. Test and Predicted labels distribution comparison

3.5 System Implementation and Deployment

Designed Online Learning System Overview

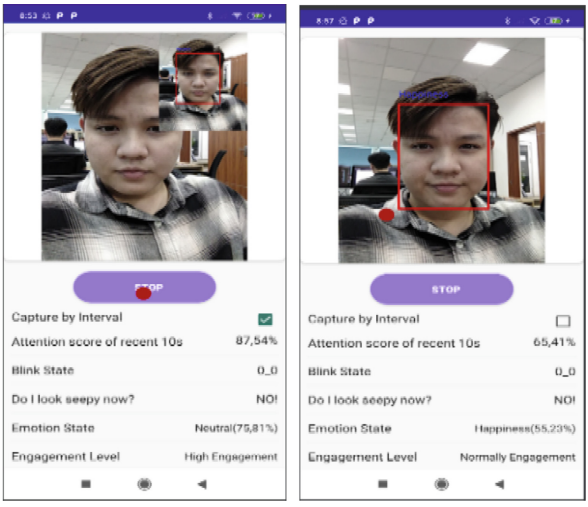

Fig. 8. Test Screen for engagement evaluation on the mobile app

This section looks into the architecture and features of the online learning system, focusing on its capability to consistently monitor student engagement and provide real-time feedback to educators. The system's core objective is to enhance the digital educational experience by fostering increased interaction and commitment between learners and educators. Our experiments target a two-part proof-of-concept system: a student-centric mobile application and an educator-centric web application, both streamlining virtual interactions.

The mobile application includes a test screen that depicts underlying operations. Over time, images of students are collected (See Fig. 8) and subsequently examined to extract metadata encompassing emotional states and engagement metrics. This metadata, once it meets specified criteria, is stored in a distant database. Post-session, educators

can access this data via the web application, where visual representations, such as charts (see Fig. 9), provide insights into the session's effectiveness and students' emotional trajectories using radar charts.

Crafting a neural network from the ground up can be both resource-intensive and time-consuming, especially when constrained by a limited dataset. Such processes often stumble upon hurdles resulting in compromised accuracy. These challenges encompass issues like data size constraints and imbalances in label distribution, which curtail the model's proficiency in feature extraction from images. Good models may underperform when paired with too datasets, underscoring the crucial nature of model fine-tuning and optimal learning rate determination. As a remedy, transfer learning emerges, offering the advantage of capitalizing on pre-existing models and consistently delivering superior performance, even with data restrictions. Essentially, transfer learning repurposes insights gleaned from pre-trained models to address the task at hand. Convolutional Neural Network (CNN) layers serve as detail extractors, processing multiple granularity levels. Transfer learning harnesses these insights for enhanced performance. The architecture of transfer learning divides into:

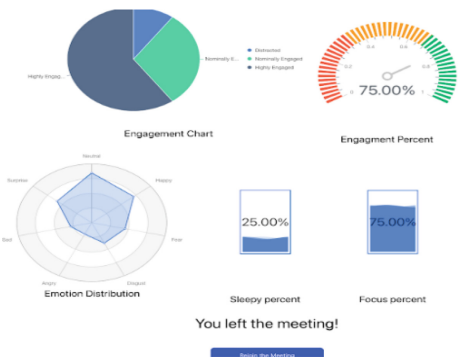

Fig. 9. Engagement metrics available after classroom session ends

- **Part 1:** The foundational network responsible for feature extraction via 2D convolutional layers. The topmost fully connected layers are omitted, assimilating this foundational network from a pre-trained model segment.
- **Part 2:** Here, fully connected layers work on data dimension reduction and compute the probability distribution for output. The primary output's unit tally mirrors the categories inherent to the classification challenge.

Comparison with Leading Approaches

The Depthwise Separable Convolution, a key feature of MobileNet, yields performance metrics on par with leading methodologies, albeit with a more compact network framework. This characteristic of MobileNet, reiterated in Tables 3, 4, & 5, demonstrates its ability in providing a lightweight model tailored for mobile device deployment [21].

Table 3. Comparing 1.0 MobileNet-224 with GoogleNet and VGG 16 on the ImageNet

Model	ImageNet Accuracy	Million Mult-Adds	Million Parameters
1.0 mobileNet-224	70.6%	569	4.2
GoogleNet	69.8%	1550	6.8
VGG 16	71.5%	15300	138

Table 4. Comparing MobileNet with Squeezenet and AlexNet

Model	ImageNet Accuracy	Million Mult-Adds	Million Parameters
0.50 mobileNet-160	60.2%	76	1.32
Squeezenet	57.5%	1700	1.25
AlexNet	57.2%	720	60

Table 5. MobileNet distilled from FaceNet

Model	1e-4 Accuracy	Million Mult-Adds	Million Parameters
FaceNet	83%	1600	7.5
1.0 MobileNet-160	79.4%	286	4.9
1.0 MobileNet-128	78.3%	185	5.5
0.75 MobileNet-128	75.2%	166	3.4
0.75 MobileNet-128	72.5%	108	3.8

3.6 Potentials for Integration with Enterprise Information System

Enterprises are continually striving to surpass their competitors, and to do so, they must harness cutting-edge technologies like machine learning, AI solutions, including NLP and sentiment extraction. Such technological advancements necessitate a modern workforce equipped with lifelong learning abilities. While these skills often lie dormant and emerge only in specific scenarios, they are challenging to measure. Despite the increasing budgets for education and training, there's a critical need for optimizing these allocations. Companies that can effectively allocate these resources gain a distinct advantage. The solution presented in this paper offers a dual advantage. First, it optimizes in-class performance by highlighting instances where learners' attention wanes, necessitating immediate intervention. Second, it allows for a continuous assessment of an individual's ability to focus and acquire new skills. This data can be incorporated into other analytic

functions, facilitating the categorization of employees based on their attention span and learning capacities.

4 Conclusion

Online education has presented numerous opportunities, yet sustaining learner engagement, especially in the financial enterprise training systems, remains an imperative challenge to address. Our research introduces a pioneering online learning tool designed for real-time attention monitoring, intertwined with emotion recognition. Using the MobileNet Convolutional Neural Network, we have successfully detected seven key emotions, translating these findings into a novel Concentration Index (CI). This CI system classifies learners into distinct engagement categories, namely "Highly engaged," "Normally Engaged," and "Disengaged."

Our model's accuracy, reflecting 70% during training and 65% during testing, affirms its robustness and efficiency. By offering these quantifiable metrics, educators are empowered with real-time feedback, enabling them to customize their teaching methods accordingly. Further, the graphical outputs serve as an insightful resource for administrators, presenting an illustrative overview of learner engagement trajectories.

In essence, this study lays down a foundational framework, propelling the enhancement of interactive and immersive virtual learning environments. As we look to the horizon, we envision harnessing broader datasets for improved emotion detection, refining the CI's accuracy, and integrating innovative technologies like chatbots. A commitment to constant evolution and embracing collaborative efforts is paramount to remain attuned to the ever-shifting landscape of online education, ensuring we continually hone engagement metrics and intervention strategies.

References

1. https://emerline.com/blog/e-learning-in-banking
2. Meyen, E.L., Aust, R.J., Bui, Y.N., Isaacson, R.: Assessing and monitoring student progress in an e-learning personnel preparation environment. Teach. Educ. Spec. Educ. **25**, 187–198 (2002)
3. Guncaga, J., Lopuchova, J., Ferdianova, V., Zacek, M., Ashimov, Y.: Survey on online learning at universities of Slovakia, Czech Republic and Kazakhstan during the COVID-19 pandemic. Educ. Sci. **12**, 458 (2022)
4. Dumford, A.D., Miller, A.L.: Online learning in higher education: exploring advantages and disadvantages for engagement. J. Comput. High. Educ. **30**, 452–465 (2018)
5. Ortony, A., Clore, G.L., Collins, A.: The Cognitive structure of emotions cambridge. UK: Cambridge University Press (1988)
6. Krithika, L.B., GG, L.P.: Student emotion recognition system (SERS) for e-learning improvement based on learner concentration metric. Procedia Comput. Sci. **85**, 767–776 (2016)
7. Picard, R.W., et al.: Affective learning—a manifesto. BT Technol. J. **22**, 253–269 (2004)

8. de Vicente, A., Pain, H.: Motivation diagnosis in intelligent tutoring systems. In: Goettl, B.P., Halff, H.M., Redfield, C.L., Shute, V.J. (eds.) Intelligent tutoring systems, pp. 86–95. Springer, Heidelberg (1998). https://doi.org/10.1007/3-540-68716-5_14
9. Du Boulay, B., Luckin, R.: Modelling human teaching tactics and strategies for tutoring systems. Int. J. Artif. Intell. Educ. **12**, 235–256 (2001)
10. de Vicente, A., Pain, H.: Informing the detection of the students' motivational state: an empirical study. In: Cerri, S.A., Gouardères, G., Paraguaçu, F. (eds.) Intelligent Tutoring Systems: 6th International Conference, ITS 2002 Biarritz, France and San Sebastian, Spain, June 2–7, 2002 Proceedings, pp. 933–943. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-47987-2_93
11. Den Uyl, M., Van Kuilenburg, H.: The FaceReader: online facial expression recognition. In: Proceedings of Measuring Behavior, pp. 589–590. Citeseer (2005)
12. Happy, S., George, A., Routray, A.: A real time facial expression classification system using local binary patterns. In: 2012 4th International Conference on Intelligent Human Computer Interaction (IHCI), pp. 1–5. IEEE (2012)
13. Woolf, B., Burelson, W., Arroyo, I.: Emotional intelligence for computer tutors. In: Workshop On Modeling and Scaffolding Affective Experiences to Impact Learning at 13th International Conference on Artificial Intelligence in Education, Los Angeles, California. (2007)
14. Feidakis, M., Daradoumis, T., Caballé, S., Conesa, J.: Measuring the Impact of Emotion Awareness on e-learning Situations. In: 2013 Seventh International Conference on Complex, Intelligent, and Software Intensive Systems, pp. 391–396. IEEE (2013)
15. Startsev, M., Zemblys, R.: Evaluating eye movement event detection: a review of the state of the art. Behav. Res. Methods **55**, 1653–1714 (2023)
16. Deubel, H., Schneider, W.X.: Saccade target selection and object recognition: evidence for a common attentional mechanism. Vision. Res. **36**, 1827–1837 (1996)
17. Lu, Y., Liu, C., Chang, F., Liu, H., Huan, H.: JHPFA-Net: Joint Head Pose and Facial Action Network for Driver Yawning Detection Across Arbitrary Poses in Videos. IEEE Trans. Intell. Transp. Syst. **24**, 11850–11863 (2023)
18. FER-2013 (2013)
19. Ezerceli, Ö., Eskil, M.T.: Convolutional neural network (CNN) algorithm based facial emotion recognition (FER) system for FER-2013 dataset. In: 2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), pp. 1–6. IEEE (2013)
20. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**, 1499–1503 (2016)
21. Howard, A.G., et al.: Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
22. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)
23. Sharma, P., et al.: Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. In: Reis, A., Barroso, J., Martins, P., Jimoyiannis, A., Huang, R.-M., Henriques, R. (eds.) Technology and Innovation in Learning, Teaching and Education: Third International Conference, TECH-EDU 2022, Lisbon, Portugal, August 31–September 2, 2022, Revised Selected Papers, pp. 52–68. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-22918-3_5
24. Mishra, S., Minh, C.S., Chuc, H.T., Long, T.V., Nguyen, T.T.: Automated Robot (Car) using Artificial Intelligence. In: 2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE), pp. 319–324. IEEE (2021)
25. https://docs.seeso.io/

26. Mishra, S., Thanh, L.T.: SATMeas - object detection and measurement: canny edge detection algorithm. In: Pan, X., Jin, T., Zhang, L.-J. (eds.) Artificial Intelligence and Mobile Services – AIMS 2022: 11th International Conference, Held as Part of the Services Conference Federation, SCF 2022, Honolulu, HI, USA, December 10–14, 2022, Proceedings, pp. 91–101. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-23504-7_7
27. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**, 1–48 (2019)
28. Tajbakhsh, N., et al.: Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imaging **35**, 1299–1312 (2016)