2023 International Conference on Emerging Smart Computing and Informatics (ESCI)
AISSMS Institute of Information Technology, Pune, India. Mar 1-3, 2023

Student Engagement Prediction in MOOCs Using Deep Learning

Naeem Ahmad
Department of MCA
NIT Raipur, India
nahmad.mca@nitrr.ac.in

Zubair Khan
Department of MCA
NIT Raipur, India
krzubairkhan@gmail.com

Deepak Singh
Department of CSE
NIT Raipur, India
dsingh.cs@nitrr.ac.in

2023 International Conference on Emerging Smart Computing and Informatics (ESCI) | 978-1-6654-7524-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ESCI56872.2023.10100007

***Abstract*—The level of peoples' engagement in a certain task is determined using an automated recognition devices (e.g. physiological sensors and pressure-sensing chairs). Even though these tools were being used in previous research works, they were very expensive and intrusive. Presently, the use of RGB video cameras is affordable and has also shown a significant effect in predicting people's engagement in tasks. Statistical tools are providing a strong foundation to model the automatic engagement identification techniques that use video cameras. In this paper, a lightweight MobileNetv2 is used to automatically determine the student engagement in MOOCs for devices with limited resources. All of the layers in the MobileNetV2 architecture have been fine-tuned to improve learning and adaptability. Instead of 1000 classes as in ImageNet, the final layer is adjusted to 3 classes of output at the final classification step. The experimental study is done on open source dataset created by subjects watch videos in online courses. Results from the evaluation phases show that our model performs better than the other two pre-trained networks (ResNet50, InceptionV4).**

***Index Terms*—Deep Learning, Student Engagement, Engagement Prediction, MOOCs, Transfer Learning**

I. Introduction

Peoples' voices, hand gestures and facial expressions are the key abilities to interact with other peoples. They also have the ability to determine the engagement level of their partners during interaction, which help in deciding the next action. A lot of efforts have been made in the domain of computer vision and its related fields to copy such abilities through intelligent systems as they are applicable in many areas such as human-robot interaction, online education, and viewers rating of videos. Previous studies [19], [24] have defined engagement in different contexts including interaction, education and technology. In this work, we focused on automatically identifying the level of engagement in an e-learning environment. Our research aims to identify the perceived engagement experienced by external observers as shown in Figure 1. The development of automatic engagement recognition systems frequently uses this perceived engagement as a target engagement [31], [8], [9], [28].

From the previous study [30] in education, it is observed that perceived engagement is one of the criteria for teachers to assess the engagement level and adjust their teaching methods accordingly. Now it is necessary to extend the use of automatic perceived engagement to online learning platforms. This automatic perceived engagement could be useful to improve

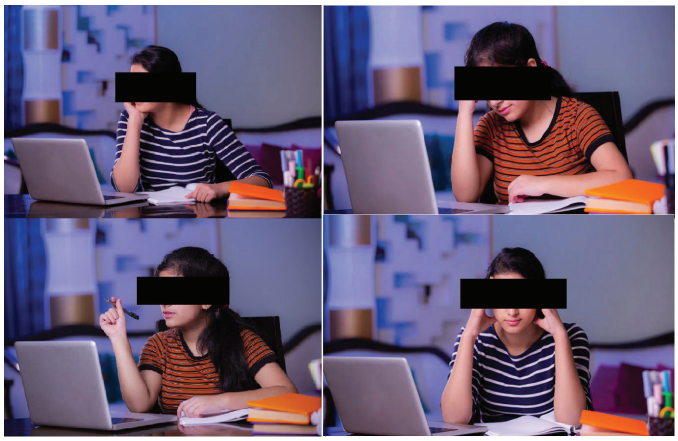

Fig. 1. Perceived Engagement in Online Classes (Images are taken from [2])

the methodology used in e-learning platforms. The definition of engagement in the context of education is based on the behavior and emotions [6], [29], which is also used in our study for perceived engagement. For example, behaviors such as being busy in other work, less attentive, talking to others, following a task could be assessed by the criteria of behavioral engagement. While students' facial expressions or less interest could be judged by emotional engagement criteria.

Cognitive engagement is difficult to recognize based on the observation alone as it is not directly associated with perceived engagement, especially in the case of students [6]. According to the study in [6], engagement has aspects of multiple sides. Hence, it is required to focus on two other types of engagement including behavioral and emotional. Apparently these three factors are dynamically associated with each other in an individual. According to the result of this study, perceived engagement is related to the behavior and emotions directly, and to cognitive engagement indirectly.

Various research efforts have been made to build automatic engagement recognition systems in previous studies [7]. However, the techniques based on RGB video camera are more often used because they are less expensive and non-intrusive [31], [9]. Following are the steps in the identification process of engagement most commonly used with video segments: first, video frames are utilized to produce a low-level feature in a specified time period. Second, the aggregation of low-

978-1-6654-7524-2/23/$31.00 ©2023 IEEE

level features obtained in previous step is used to extract a high-level feature. Finally, these high-level features are used in decision making. However, The said process using feature learning of data-driven is not popular due to insufficient data for data-driven machine learning approaches.

Recently, deep learning has gradually got more popularity in overcoming problems of conventional machine learning in facial images to determine the level of students' engagement. However, this popularity of deep learning is not possible without the use of enough computational resources and a massive amount of data. It is seen that these three factors are not always available while using edge devices due to low computational power and less memory. Therefore, a light-weight deep learning model is needed to overcome these limitations [21]. Conventional methods based on CNNs fail to achieve high accuracy as they require large datasets. In order to overcome these limitations, many improvisations of CNN models have been proposed [16]. In past decade, research efforts focused on designing lightweight architecture without compromising performance for facial images [14]. Such types of lightweight models would work on the devices with limited resources, which would be greatly helpful for academicians. One possible solution would be a lightweight CNNs model that would classify facial images [32].

One of the common problems with the data is class imbalance which requires an effective classification approach. Although several studies have presented the improvised deep learning, a very limited work is done for problem of imbalanced data in deep learning. These existing techniques can be broadly classified based on the working of algorithm and data, and their ensembling. In this paper, a fine-tuned lightweight model is proposed to deal with class imbalance of the dataset. Here, data augmentation is applied on the training and classified imbalanced dataset using pre-trained MobileNetv2 model. The main objective of this paper is to propose an automated recognition method based on deep learning for determining the level of students' engagement using facial image dataset.

The rest of papers is organised as follows: We discuss related work that has already been done in Section II in the area of engagement detection, perceived engagement and other related researches. The proposed methodology for engagement recognition is described in Section III. The obtained results are discussed in Section IV. Finally, We conclude our work in Section V.

II. Related Works

Various researcher studies in human-to-any interaction (may be human for HHI, computer for HCI, and resources for (HRI)) technology have investigated the meaning of engagement [24]. In [17], Authors have explained that engagement is a mental state that includes the point of engagement, the period of sustained engagement, the point of disengagement, and the point of re-engagement. Although there is no universally agreed-upon definition of engagement, authors focused on the theme of maintaining and continuing participation between the parties involved. Research studies on engagement in HRI and HCI can be divided into two distinct groups. The first viewpoint is that of figuring out how to make the robot more interesting to interact with the participant. The other viewpoint is automatic identification of human engagement during interaction [15].

Our research is more in line with the second point of view. Our research focused on developing a system that could detect student interest and participation in a learning environment automatically. Students' participation level is described in several context by the educational researcher community. In literature [6], study presented 3 types of engagement comprising cognitive, emotional, and behavioural. Various scholars have proposed other forms of involvement, such as behavioural, intellectual, cognitive, and psychological [1]. Prior studies [9], [13] on automatic engagement detection focused on the perceived engagement. As the students' feedback on delivery content of an instructor is based on the perceived engagement, our study of automatic engagement detection also utilize the perceived engagement.

Several automatic engagement detection system have been proposed using different types input signals [31], [4], [18], [12]. Speech, physiological, visual, contextual, and multimodal data are all examples of input signals. We focused on the literatures that were most relevant to our research. The authors in [31] developed an automatic detection system to categorise students' levels of engagement while working on a computer-based educational puzzle. With the use of a camera and a pressure-sensitive chair, they were able to analyse facial expressions and head movements. Their research was the first attempt to measure an individual's interest.

An algorithm was proposed to determine the level of interest in a meeting situations. The system used a Hidden Markov Model-based recognition technique to identify the high-interest subset of the group based on audio-visual inputs. Their research showed that auditory information was more important than visual [20]. In 2015, authors presented a study on the impact of varying group size by comparing categorization models trained on data collected from individuals and from groups in HRI [15]. The research states that models trained using data on group interactions can be used to a wider variety of scenarios. They used both manually and automatically retrieved features from audio, visual, and contextual data to detect disengagement. The authors in [28] introduced a mechanism for automatically providing feedback to teachers that help in determining the level of student participation. Students' gaze points were determined using data on their heads in motion. They used students' gaze data as a proxy for their level of participation in class.

Authors in [31] developed a system that analyses a sequence of images to determine a student's level of interest based solely on their facial expressions. They demonstrated the feasibility of using two engagement levels and a video clip of 10 seconds is most discriminative length. Also some methods were developed for accumulating features or cues from the frames of an image given in a specific time period. When modelling

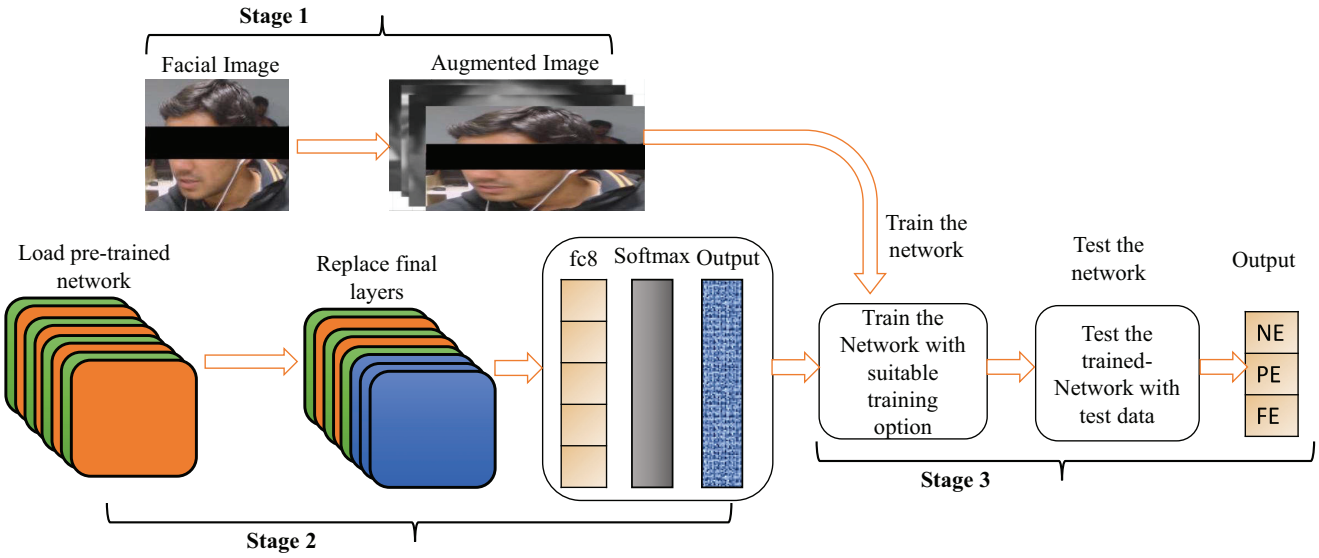

Fig. 2. Workflow of multi-classification based on transfer learning

temporal dynamics, statistical functionals are typically utilised as an aggregation strategy [8], [21], [27], [23]. Authors in [27] proposed a model that uses analysis of human postures and body motion to predict engagement levels of children. In their model, temporal series features are aggregated and transformed into meta features using min, max, mean, and normalised histogram functions.

In order to classify videos in a deep learning framework, the authors devised an approach that combines data from single-frame CNNs [26]. For time information fusion, they looked into CNNs with multiple connections in the temporal domain and came up with 3 different fusion models including Late, Early, and Slow. The proposed models are different in how they combine this information in a time domain. The authors in [11] demonstrated that the results from the Slow Fusion model are superior to those from the Late Fusion model and to those from the Early Fusion model. Several neural networks learn spatial and temporal features using 3D convolutional kernels [10]. Our method is related to the MobileNetV2 model [22]. Using a 3D convolutional layer, this model collects each spatial feature map from the input images. It is seen that freezing the pre-trained network in low layers has been shown to improve accuracy more than fine-tuning it [5]. With this literature survey, it is fond that a lightweight CNNs model is needed to classify facial image and predict the student engagement level in devices with limited resources which would be greatly helpful for academicians.

III. Proposed Methodology

CNNs are found to be the best models for processing and analysing the image data as it has shown the outstanding performance in image segmentation, image classification, and many other applications. CNNs is a feed-forward network having multi-level hierarchical arrangement of many layers. The basic configuration of CNNs is alternate layers of convolution and pooling coupled to fully connected layers at the endpoint. The collections of convolution kernels are made at each layer that does multiple transformations. Sometimes, global average pooling layer is used to restored the fully connected layer [3], [14]. The CNNs based deep learning models and mobile devices are widely used for various industrial applications. Mobile devices have limited resources and computational power, which require lightweight deep learning models. Determining the local areas in an image using convolution extraction is primary operation of CNNs. However, this operation suffers with two main problems: (1) lack of direct global features extraction and (2) complex computation due to increase in convolution kernel parameters. Different convolution methods have been proposed to address these problems. Dilated convolution, group convolution, depth-wise separable convolution and deformable convolution are very famous methods [32]. Here, a pre-trained networks model MobileNetV2 with transfer learning is applied, which utilizes the depth-wise separable convolutions.

A. MobileNetV2

MobileNetV2 is a pre-trained networks which belongs the class of lightweight models. It uses an inverted residual structure, linear bottlenecks and lightweight depth-wise separable convolution filters. These filters operate at the middle layers and contribute to non-linearity. This entire architecture has been specifically built for mobile devices having limited resources and computational power [22]. The primary building blocks of this architecture are of two types: (1) residual block with stride and (2) downsizing with stride 2. It comprises layers including Conv 1x1 with ReLU6, depth-wise separable

convolution, and Conv 1x1 without any non-linearity. Hence, overall architecture of MobileNetV2 contains 53 layers.

B. Transfer Learning Paradigm

Transfer learning is playing vital role in improving the performance of target learners by transferring the knowledge received from different and related sources. It is required when data becomes easily outdated. It means labeled data collected in specific time window does not follow the same distribution in future time window. For example, it is possible that using a model trained on a dataset of one type of person's images in a specific time period will degrade the performance of an expression estimate when applied to different types of people in different time periods. Transfer learning utilizes the pre-trained networks fine-tuned over a new task for better learning [25]. Figure 2 depicts the classification process of students based on their engagement level using transfer learning and pre-trained MobileNetV2 network.

It is done using the following steps:

1) Create facial images from the input videos
2) Image augmentation is applied on the dataset as it contains both types of images:colored images and black & white images.
3) The low-level features are generated from the lower layers of pretrained network including MobileNetV2, ResNet-50, and IceptionV4.
4) Low-level features are passed to the upper layers to extract high-level features from the augmented images.
5) Dataset of facial images is categorized into 3 classes for training images.
6) Trained networks are tuned on the validation data and tested on tuned models for the test images.

In this work, all layers of MobileNetV2 is fine-tuned to perform the classification task. In order to perform comparative analysis, two more pre-trained networks (ResNet-50, IceptionV4) are used where each layer is fine-tuned using the transfer learning to perform the classification task.

IV. Results and Discussion

In this section, obtained results are presented for comparative analysis. This comparison is done between lightweight MobileNetv2 model and two other models including ResNet-50, IceptionV4. In order to show the comparison, We have used open source data downloaded from https://github.com/e-drishti/wacv2016. This data is captured for subjects watch videos in online courses and labelled using crowdsourcing with varying option of engagement labels. Because the dataset contains three (Not-engaged, Partially-engaged, Very-engaged) classes and the input image size is 100x100x3, a wide variety of experimental configurations were considered. In training phase, 70% of the input images were used, while 30% of the input images were used for testing. The training parameters are a batch size of 16, gamma value of 0.1 for an Adam solver, a learning rate of 0.001, and a step size of 7. The model is trained for 25 epochs. The model is trained using a Cross-Entropy loss. All of the layers in the MobileNetV2 architecture have been fine-tuned. Instead of 1000 classes like ImageNet, the final dense layer is adjusted to output three classes.

Deep Learning Toolbox of MATLAB has been used for all experiments associated with this study. Three said pre-trained networks are trained and tested on the e-learning dataset. These model are fine-tuned by modifying the output of last layer to three classes (Very-engaged, Nominally-engaged and Not-engaged) to perform the classification task. The results show that lightweight MobileNetv2 model is useful for devices with limited resources instead of heavy deep CNNs for classification task without compromising performance. Average accuracy of the proposed model is 74.55% better than other two pre-trained networks given in Table I. Training validation accuracy, and loss plots of proposed model is shown in Figures 3 and 4 respectively.

V. Conclusion

This work presents an automated recognition system of the students' engagement in online sessions using lightweight pre-trained networks, which is useful for devices with limited resources. Here, a multi-class (Very-engaged, Nominally-engaged, Not-engaged) is performed using crowdsource e-learning database. Transfer learning-based lightweight model is developed which is obtained by fine tuning the layers of MobileNetv2. Obtained results are compared with two other pre-trained networks ResNet-50 and IceptionV4. Lightweight MobileNetv2 outperforms with average accuracy of 74.55%. The results show that lightweight MobileNetv2 model is useful for devices with limited resources instead of heavy deep CNNs for classification task without compromising performance. In future, this work can be extended to any advanced deep learning methods for edge devices, which may be helpful for academicians and instructors in online classes.

Acknowledgement

This work is part of major project of MCA students supported by Department of Computer Applications, National Institute of Technology (NIT) Raipur, India.

References

[1] M. E. Alvarez and A. J. Frey, "Promoting academic success through student engagement," pp. 1–2, 2012.

[2] S. Anand, "Are online classes wearing children out?" in *India Today Insight*, November 23,2020.

[3] R. K. Barbhuiya, N. Ahmad, and W. Akram, "Application of convolutional neural networks in cancer diagnosis," in *Computational Intelligence in Oncology, Springer*, 2022, pp. 95–109.

[4] M. Dewan, M. Murshed, and F. Lin, "Engagement detection in online learning: a review," *Smart Learning Environments*, vol. 6, no. 1, pp. 1–20, 2019.

[5] C. Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 203–213.

[6] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "School engagement: Potential of the concept, state of the evidence," *Review of educational research*, vol. 74, no. 1, pp. 59–109, 2004.

[7] A. Ghandeharioun, D. McDuff, M. Czerwinski, and K. Rowan, "Emma: an emotion-aware wellbeing chatbot," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 1–7.

TABLE I
DIFFERENT PRE-TRAINED NETWORKS USED FOR THE STUDY

Folds	Pre-trained	Accuracy	Sensitivity	Specificity	Precision
Fold-1	Inception-V4	0.7159	0.7351	0.8579	0.6563
	Resnet 50	0.7186	0.7505	0.8546	0.6622
	Mobilenetv2	0.7371	0.7395	0.8866	0.6860
	Proposed	0.7451	0.7457	0.8936	0.6920
Fold-2	Inception-V4	0.7005	0.7469	0.8514	0.6498
	Resnet 50	0.7125	0.7528	0.8618	0.6623
	Mobilenetv2	0.7344	0.7468	0.8793	0.6798
	Proposed	0.7424	0.7498	0.8854	0.6891
Fold-3	Inception-V4	0.7086	0.7359	0.8466	0.6534
	Resnet 50	0.7129	0.7338	0.8693	0.6644
	Mobilenetv2	0.7318	0.7504	0.8679	0.6801
	Proposed	0.7461	0.7617	0.8734	0.6967
Fold-4	Inception-V4	0.7107	0.7440	0.8523	0.6562
	Resnet 50	0.7175	0.7404	0.8615	0.6597
	Mobilenetv2	0.7387	0.7355	0.8665	0.6747
	Proposed	0.7462	0.7487	0.8746	0.6882
Fold-5	Inception-V4	0.7184	0.7339	0.8713	0.6674
	Resnet 50	0.7156	0.7525	0.8607	0.6625
	Mobilenetv2	0.7389	0.7762	0.8667	0.6836
	Proposed	0.7477	0.7887	0.8773	0.6976

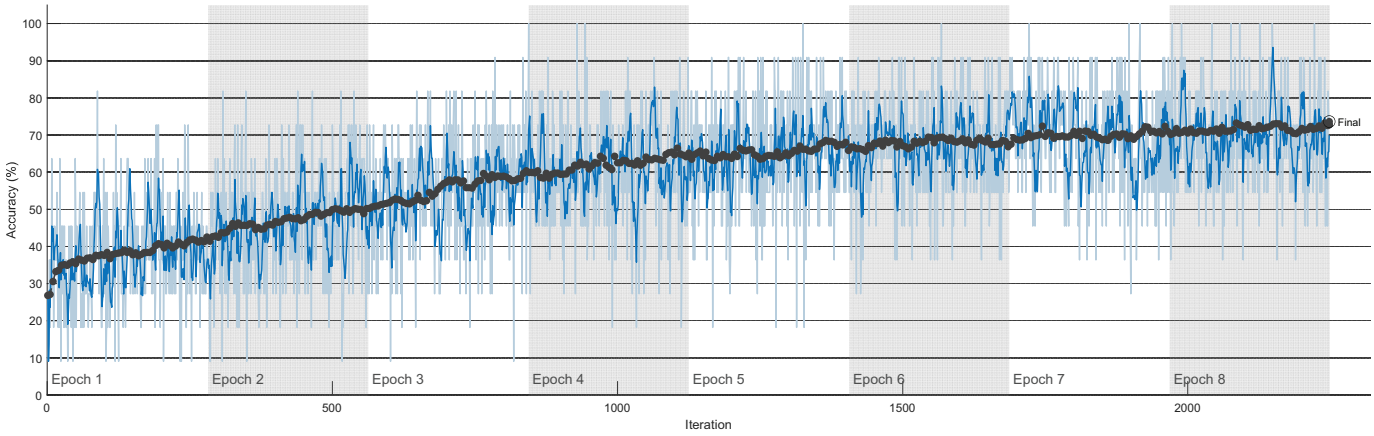

Fig. 3. Accuracy plot of proposed model

[8] J. Hernandez, Z. Liu, G. Hulten, D. DeBarr, K. Krum, and Z. Zhang, "Measuring the engagement level of tv viewers," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 2013, pp. 1–7.

[9] M. Jang, D.-H. Lee, J. Kim, and Y. Cho, "Identifying principal social signals in private student-teacher interactions for robot-enhanced education," in *2013 IEEE RO-MAN*. IEEE, 2013, pp. 621–626.

[10] C. Jing, P. Wei, H. Sun, and N. Zheng, "Spatiotemporal neural networks for action recognition based on joint loss," *Neural Computing and Applications*, vol. 32, pp. 4293–4302, 2020.

[11] A. Kamel, B. Sheng, P. Yang, P. Li, R. Shen, and D. D. Feng, "Deep convolutional neural networks for human action recognition using depth maps and postures," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 9, pp. 1806–1819, 2018.

[12] A. Khan, J. P. Li, N. Ahmad, S. Sethi, A. U. Haq, S. H. Patel, and S. Rahim, "Predicting emerging trends on social media by modeling it as temporal bipartite networks," *IEEE Access*, vol. 8, pp. 39 635–39 646, 2020.

[13] A. Khan, J. P. Li, A. u. Haq, S. Nazir, N. Ahmad, N. Varish, A. Malik, and S. H. Patel, "Partial observer decision process model for crane-robot action," *Scientific Programming*, vol. 2020, pp. 1–14, 2020.

[14] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial intelligence review*, vol. 53, no. 8, pp. 5455–5516, 2020.

[15] I. Leite, M. McCoy, D. Ullman, N. Salomons, and B. Scassellati, "Comparing models of disengagement in individual and group interactions," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2015, pp. 99–105.

[16] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems*, 2021.

[17] H. L. O'Brien and E. G. Toms, "What is user engagement? a conceptual framework for defining user engagement with technology," *Journal of the American society for Information Science and Technology*, vol. 59, no. 6, pp. 938–955, 2008.

[18] C. Oertel, G. Castellano, M. Chetouani, J. Nasir, M. Obaid, C. Pelachaud, and C. Peters, "Engagement in human-agent interaction: An overview," *Frontiers in Robotics and AI*, vol. 7, p. 92, 2020.

[19] C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, and I. Poggi, "A model of attention and interest using gaze behavior," in *International Workshop on Intelligent Virtual Agents*. Springer, 2005, pp. 229–240.

[20] A. Plopski, T. Hirzle, N. Norouzi, L. Qian, G. Bruder, and T. Langlotz, "The eye in extended reality: A survey on gaze interaction and eye tracking in head-worn extended reality," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–39, 2022.

[21] M. S. Ryoo, B. Rothrock, and L. Matthies, "Pooled motion features for first-person videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 896–904.

[22] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings*

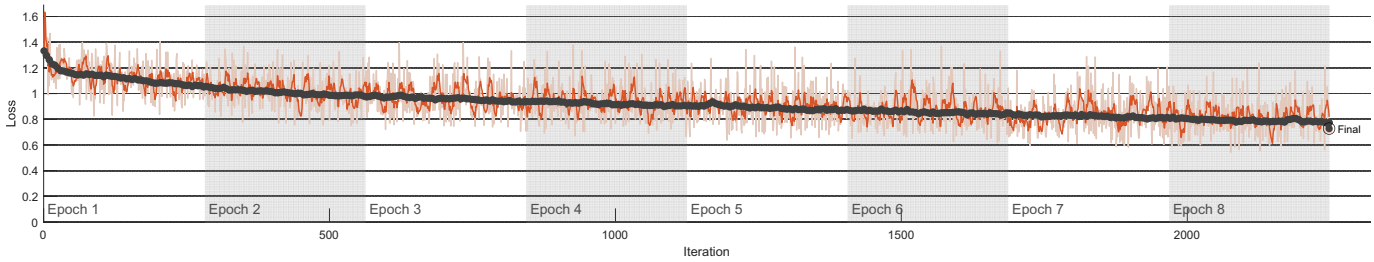

Fig. 4. Loss plot of proposed model

of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.

[23] S. Senecal, L. Cuel, A. Aristidou, and N. Magnenat-Thalmann, "Continuous body emotion recognition system during theater performances," *Computer Animation and Virtual Worlds*, vol. 27, no. 3-4, pp. 311–320, 2016.

[24] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, no. 1-2, pp. 140–164, 2005.

[25] D. Singh, A. Shukla, and M. Sajwan, "Deep transfer learning framework for the identification of malicious activities to combat cyberattack," *Future Generation Computer Systems*, vol. 125, pp. 687–697, 2021.

[26] B. SravyaPranati, D. Suma, C. ManjuLatha, and S. Putheti, "Large-scale video classification with convolutional neural networks," in *Information and Communication Technology for Intelligent Systems: Proceedings of ICTIS 2020, Volume 2*. Springer, 2021, pp. 689–695.

[27] B. Stephens-Fripp, F. Naghdy, D. Stirling, and G. Naghdy, "Automatic affect perception based on body gait and posture: A survey," *International Journal of Social Robotics*, vol. 9, no. 5, pp. 617–641, 2017.

[28] Ö. Sümer, P. Goldberg, S. D'Mello, P. Gerjets, U. Trautwein, and E. Kasneci, "Multimodal engagement analysis from facial videos in the classroom," *IEEE Transactions on Affective Computing*, 2021.

[29] P. Sunitha, N. Ahmad, and R. K. Barbhuiya, "Impact of covid-19 on education," in *ICCCE 2021,Springer*, 2022, pp. 1191–1197.

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[31] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, 2014.

[32] Y. Zhou, S. Chen, Y. Wang, and W. Huan, "Review of research on lightweight convolutional neural networks," in *2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)*. IEEE, 2020, pp. 1713–1720.