Automated Classification of Student's Emotion Through Facial Expressions Using Transfer Learning

Rajamanickam Yuvaraj, Nanyang Technological University, Singapore
Ratnavel Rajalakshmi, Vellore Institute of Technology, India
Venkata Dhanvanth, Vellore Institute of Technology, India
Jack Fogarty, Nanyang Technological University, Singapore

The European Conference on Education 2023
Official Conference Proceedings

Abstract

Emotions play a critical role in learning. Having a good understanding of student emotions during class is important for students and teachers to improve their teaching and learning experiences. For instance, analyzing students' emotions during learning can provide teachers with feedback regarding student engagement, enabling teachers to make pedagogical decisions to enhance student learning. This information may also provide students with valuable feedback for improved emotion regulation in learning contexts. In practice, it is not easy for teachers to monitor all students while teaching. In this paper, we propose an automated framework for emotional classification through students' facial expression and recognizing academic affective states, including amusement, anger, boredom, confusion, engagement, interest, relief, sadness, and surprise. The methodology includes dataset construction, pre-processing, and deep convolutional neural network (CNN) framework based on VGG-19 (pre-trained and configured) as a feature extractor and multi-layer perceptron (MLP) as a classier. To evaluate the performance, we created a dataset of the aforementioned facial expressions from three publicly available datasets that link academic emotions: DAiSEE, Raf-DB, and EmotioNet, as well as classroom videos from the internet. The configured VGG-19 CNN system yields a mean classification accuracy, sensitivity, and specificity of 82.73% ± 2.26, 82.55% ± 2.14, and 97.67% ± 0.45, respectively when estimated by 5-fold cross validation. The result shows that the proposed framework can effectively classify student emotions in class and may provide a useful tool to assist teachers understand the emotional climate in their class, thus enabling them to make more informed pedagogical decisions to improve student learning experiences.

Keywords: Classroom, Students, Facial Expressions, Education, Machine Learning

iafor
The International Academic Forum
www.iafor.org

Introduction

Emotions play a major role in guiding cognition and behavior. In education or learning contexts, emotions can affect students' attention, motivation (Skinner, Pitzer, & Brule, 2014), and their use of learning strategies (Pekrun, 2014). Measuring emotions in educational settings can also provide important information explaining (and possibly predicting) students' learning outcomes. For example, students' emotions in class can reflect their feelings about course content and indicate their engagement, which has been found to predict academic performance (Reyes, Brackett, Rivers, White, & Salovey, 2012). Feedback regarding students' emotional responses during class may also be used to develop and optimize the learner's experience. Hence, it is of great value to explore the emotions of students as they learn. Students' facial expressions can convey their comprehension of information and their emotions during learning and experienced educators can often adjust their teaching according to students' expressions. However, it can be difficult to monitor students continuously, particularly for less experienced teachers. Objective observational data is also needed to study and reflect on the role of emotions in different learning contexts, and for this to be developed into an accessible tool that educators may be able to use to improve teaching practice. Machine learning (ML) offers a promising method for monitoring student emotions by classifying facial expressions from video data (Hu et al., 2020; Zeng et al., 2020; Yuan, 2022). The present study builds on that by investigating an automatic classification system using ML to categorize students' emotions through facial expressions during class.

Several studies have investigated ML-based emotion recognition using video data recorded in real classes; however, those studies have mainly focused on classifying Ekman's basic emotions, including happiness, sadness, fear, anger, disgust, and surprise (Zeng et al., 2020; Yadegaridehkordi, Noor, Ayub, Affal, & Hussin, 2019). Few studies have aimed to recognize (classify) expressions of academic emotions, such as boredom, engagement, confusion, frustration, surprise, relief, sadness, and joy (Hu et al., 2020; Yuan, 2022; Gupta, D'Cunha, Awasthi, & Balasubramanian, 2016). Academic emotions are defined as emotions of a student experienced in academic settings such as class-related or learning related situations and have been associated to academic performance, motivation to learn, learning strategies, and self-regulation (Pekrun, Goetz, Frenzel, Barchfeld, & Perry, 2011; Pekrun, Goetz, Titz, & Perry, 2002). Considering the relevance of academic emotions in educational contexts, and the expansive literature exploring academic emotions relative to learning outcomes, it is important to develop methods for recognizing these emotions during learning, as opposed to recognizing the six basic emotions defined by Ekman. The present study aims to develop an effective ML algorithm for classifying academic emotions, adding to the limited research in this area.

The utility of several ML algorithms has been studied for video data classification, and deep learning, in particular, has been shown to greatly improve classification accuracy relative to other methods. For that reason, an increasing number of researchers are using deep learning for the classification of emotional expressions (Zeng et al., 2020; Yuan, 2022; Gupta et al., 2016; Pabba & Kumar, 2022). Moreover, of the various deep learning models available, Convolutional Neural Networks (CNN) is often found to provide strong classification performance in various applications (Bajpai, Yuvaraj, & Prince, 2021; Thomas et al., 2020; Oh et al., 2020). Regarding emotion recognition, Yuan (2022) proposed a CNN-based system for classifying students' emotions using videos recorded during class; emotions including focus, puzzled, distracted, silence, nervous, joy, exhausted, and bored were classified with an accuracy of approximately 78.3% (Yuan, 2022). Zeng et al. (2021) also applied CNN to classify anger, surprise, happiness, neutral sadness, disgust, and fear in children and adults during learning, with accuracies reaching 64.8% and 68.5%, respectively (Zeng et al., 2020). Ashwin et al. (2019) also used a novel CNN model to classify student engagement levels with an accuracy of about 71% (Ashwin & Guddeti, 2020). The results across these studies show that CNN can provide reasonable classification performance. However, there are still limitations associated with CNN-based methods for face recognition: First, training an effective CNN model requires a large number of labeled facial images. Second, training a CNN model from scratch can be very time-consuming and computationally expensive.

Transfer learning has emerged as an effective approach to overcome the above-mentioned limitations and involves the process of taking a pretrained deep learning network and fine-tuning it to learn a new task (Raghu, Sriraam, Temel, Rao, & Kubben, 2020). Using this approach, a pretrained CNN model could leverage knowledge gained from a large dataset to classify similar information in a smaller dataset more efficiently. In other words, instead of training the CNN model from scratch, fine-tuning a pre-trained CNN model for a novel dataset can significantly reduce the training time and save computational resources. Several effective real-world examples of this technique are available, including pre-trained CNN models such as VGGNet, GoogleNet, ResNet, and AlexNet (Guo et al., 2016).

In this study, a transfer CNN framework based on VGG-19 is investigated for classifying student's facial expressions of academic emotions in classroom videos. The proposed framework consists of a pre-trained VGG-19 and configured VGG-19 model used for emotion classification. Emotional expressions were categorized according to the nine academic emotions, including amusement, anger, boredom, confusion, engagement, interest, relief, sadness, and surprise, which are pertinent in typical classroom environments and education research. To that end, we first constructed a broad dataset of facial expressions representing those emotions from three publicly available datasets (i.e., DAiSEE, Raf-DB, and EmotioNet) that were previously coded for certain academic emotions, as well as other classroom videos from the internet. The VGG-19 CNN model was then pretrained on ImageNet, before configuring its structure and parameters for application to this work. The configured CNN model was then fine-tuned for emotion classification of the constructed dataset.

The main contributions of the current study are the following: i) the construction of a broad dataset with coded academic emotions, and ii) the development of a deep transfer CNN framework for academic emotion classification, which can improve and accelerate future work utilizing this approach for emotion recognition. The rest of the paper is organized as follows: Section 2 briefly introduces the dataset construction and methodological details of this study, including data pre-processing and the development of the CNN framework. Section 3 presents the study results and discussion, and Section 4 provides a general conclusion, with limitations and directions for future work.

Materials and Methods

Figure 1 shows the methodological framework used in the present study, including dataset construction, video data preprocessing, CNN framework development, and emotion classification. All the analysis are carried out with the Kera's deep learning framework on a high-performance computer, which is equipped with an Intel Core i5-8265U CPU @1.60GHz (8 CPUs) 1.8GHz, 8 Gigabyte (GB) RAM, and 4 GB NVIDIA GeForce MX250 graphics card.

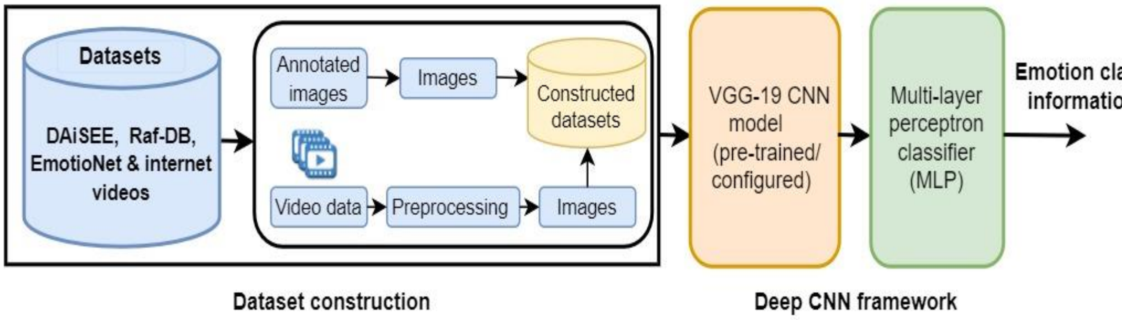
Datasets
DAiSEE, Raf-DB, EmotioNet & internet videos
Annotated images
Images
Constructed datasets
Video data
Preprocessing
Images
VGG-19 CNN model (pre-trained/ configured)
Multi-layer perceptron classifier (MLP)
Emotion class information
Dataset construction
Deep CNN framework

Figure 1: An overview of the proposed framework for classroom emotional climate classification based on students' facial expressions.

Dataset Construction

Dataset	Description	Dur.	Emotions	Environment
DAISEE	3935 images annotated from 9068 videos	10	Boredom, confusion, engagement, frustration	E-learning
Raf-DB	29,672 images annotated queried from the web	NA	Surprise, fear, joy, disgust, anger, sadness	-
EmotioNet	100,000 images annotated queried from the web	NA	23 emotions	In wild
Classroom videos from the Internet	1812 annotated images from 56 videos	10 to 30	Relief, amusement	In classroom

Table 1: Information about the datasets used in this study. NA-not applicable; "-"means that this information not available in the dataset. Dur.-Duration; Duration in seconds.

A comprehensive dataset of students' facial expressions reflecting the nine academic emotions (Pekrun et al., 2002) defined by Pekrun et al. (2002) was constructed using video data from multiple sources, including three publicly available and coded datasets (i.e., DAiSEE (Gupta et al., 2016), Raf-DB (Li, Deng, & Du, 2017), EmotioNet (Fabian Benitez-Quiroz, Srinivasan, & Martinez, 2016)), and other classroom videos available on the internet. Classroom videos were selected by query-based method using the key words 'classroom videos', 'classroom teaching', and 'students in the classroom' from the website https://www.istockphoto.com/. A summary of all sources used for the construction of this larger dataset is listed in Table 1. This study relies on static facial images (color) for recognizing students' emotion. Whereas some datasets listed in Table 1 contained video clips. By applying pre-processing techniques, we extracted static facial images from the video clips, explained in Section 2.2. Finally, a balanced hybrid dataset was constructed for all 9 academic facial expressions in a total of 8674 colour images of size 128 × 128 pixels were utilized. Table 2 provides details of the contribution of various sources in the dataset construction for this study. This constructed dataset is utilized for CNN model architecture to classify emotions during class.

Preprocessing

The classroom video clips undergo a set of pre-processing steps including frame sampling, face detection, extraction, and resizing. Each video contains a large number of frames and frames in close proximity are almost the same. According to a study by (Pabba & Kumar, 2022), the result obtained by processing four video frames/sec with a time interval of 0.25 sec is almost equal to the result which is obtained by processing 30 video frames per second. Therefore, in this frame sampling step, only four video frames are processed per second with a time interval of 0.25 seconds, thereby reducing computational overhead. Given the proximity of students to one another, many frames included several students; hence, it was necessary to detect and extract individual students' faces from each frame. To achieve that, a deep learning model called PyFeat was applied to detect individual faces in each sampling frame, before cropping the detected face form the whole image. All face images were then resized to 128 × 128 (width × height) pixels and then input to the CNN for training and evaluation.

Emotion	DAISEE	Rad-DB	EmotioNet	Int. videos	Total
Anger	-	705	222	-	927
Amused	-	-	-	1000	1000
Boredom	1000	-	-	-	1000
Confusion	1000	-	-	-	1000
Engagement	1000	-	-	-	1000
Interest	-	-	-	935	935
Surprise	-	700	300	-	1000
Relief	-	-	-	812	812
Sadness	-	1000	-	-	1000
Total	**3935**	**2405**	**522**	**1812**	**8674**

Table 2: Details of constructed dataset in this study. "-" means that this emotion is not listed in the dataset. The bold numbers represent the total number of images for each class across all datasets (rows), the total number of images from each dataset (column). Int.-Internet

CNN Framework for Emotion Classification

A transfer learning CNN framework was proposed to classify emotions from facial images. The architecture of the proposed framework is shown in Figure 2 and consists of a pre-trained VGG-19 CNN model and a configured CNN model, where the pre-trained VGG-19 CNN model is used to extract universal (generalizable) features for common image classifications tasks, and the configured CNN model is used to classify emotions from facial images.

Specific information about the pre-trained VGG-19 CNN, the configured CNN, and the training procedure of the proposed framework is detailed in the following paragraphs.

Pre-trained VGG-19 CNN: The deep neural network VGG-19 is a well-known CNN model with 19 layers, and it has achieved remarkable performance in various image processing tasks

(Raghu et al., 2020; Xu et al., 2019). VGG-19 replaces large-sized convolution filters with small-sized filters while increasing the depth of network. This is mainly because CNN with small filters will benefit the improvement of classification accuracy. Figure 2 shows the detailed configurations of all layers in VGG-19. The VGG-19 CNN model used in this paper is pre-trained on the ImageNet dataset, and the front-layers of the pre-trained CNN model can extract low-level universal features, which are appropriate for general image processing tasks.

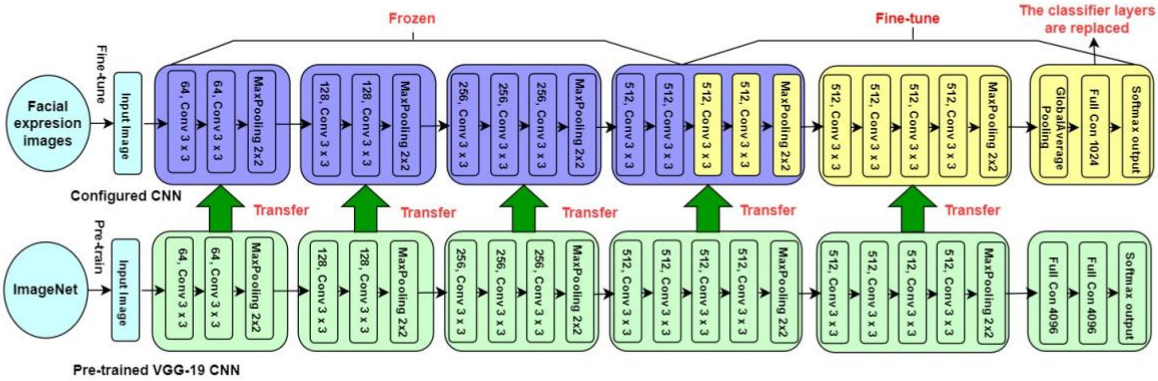

Figure 2: The proposed deep transfer CNN framework based on VGG-19

Configured VGG-19 CNN: Here, the original output layer is removed and then a new SoftMax layer is added, and it is used for emotion classification, there are nine 'neurons' in the new output matching the number of academic emotions. The hyper-parameters, parameters and structures of the pretrained CNN model are directly transferred to the configured CNN model to improve its classification performance. Then, the configured CNN model can be fine-tuned for the specific emotion classification without the need to train the whole network from scratch. The loss function for fine-tuning the configured CNN is categorical cross entropy and the epoch size is 50. In addition, we randomly drop 40% of the weights of the fully connected layer at each training iteration, to prevent overfitting on the training data and enhance generalizability. The adaptive moment estimation (Adam) optimizer (learning rate = 0.0001) was used for the training of the configured VGG-19 CNN model.

Training Procedure

After the structures of the pre-trained VGG-19 CNN and the configured CNN are successfully designed, the constructed dataset images were divided into 5-folds randomly. For each iteration step, the CNN model was trained on 4-folds to fine tune the configured CNN and we performed the testing on the remining fold (5th fold). This process is repeated five times until all the folds are used as a test set. As shown in Figure 3, the front- layers from layer 1 to layer 12 of the configured CNN are frozen. While the later layers after Layer 12 are set to be trainable, these layers are fine-tuned on the constructed dataset. The performance of the proposed framework is evaluated with the three parameters namely accuracy, sensitivity, and specificity, which are most commonly considered in the emotion recognition literature (Zeng et al., 2020; Yuan, 2022). The final results were obtained by averaging accuracies, sensitivities, and specificities of five folds.

Experimental Results and Discussion

Prominent CNN models like VGG-19 have tens of millions of parameters. If all those parameters are trained from scratch, millions of images would be needed to ensure that the network could select features properly. The demand for so many images could be almost impossible to meet when developing models for application in specific real-world contexts. However, considering images from ImageNet and our custom dataset images have common low-level features, it is possible to transfer parameters pretrained on ImageNet to our model designed for use in classroom contexts. This transfer approach was tested in the present study to develop a configured emotion recognition model that may be used to classify student emotions and classroom emotional climate. The configured model was more effective, relative to the pre-trained model, illustrating the successful application of this technique, which may be valuable in the efficient development of deep learning emotion recognition models in future research and applications.

Model	Fold	Accuracy (%)	Sensitivity (%)	Specificity (%)
Pre-trained VGG-19	1	71.07	69.28	96.02
	2	74.38	74.19	97.37
	3	70.02	71.05	96.08
	4	74.62	73.67	97.85
	5	71.72	70.83	96.61
	Average	**72.12±1.89%**	**71.85±2.21%**	**96.54±0.62%**
Configured VGG-19	1	81.36	84.23	97.42
	2	83.67	84.01	97.65
	3	83.89	85.904	97.78
	4	79.11	79.98	97.38
	5	85.62	87.37	98.21
	Average	**82.73±2.26%**	**82.55±2.14%**	**97.67±0.45%**

Table 3: Classification Performance of the pre-trained and configured VGG-19 CNN

Table 3 presents the performance of the proposed machine learning framework, including the pre-trained and con- figured VGG-19. As expected, the configured framework was more effective, showing higher accuracy, sensitivity, and specificity relative to the pre-trained model. The most notable improvements were in average accuracy and sensitivity, which increased by almost 11% after configuration. The configured VGG-19 model also delivered performance with the lowest SD of accuracy, showing greater consistency relative to the pre-trained model. Moreover, as shown in Figure 3, the superiority of the configured model is also evident when looking at performance for each facial expression category. Together, these results demonstrate the significant positive effect that transfer learning can have on CNN models, supporting the application of this technique to efficiently develop state-of-the-art deep learning models for specific classification purposes.

Table 4 shows the average computational time for each module block to process 100 faces using the pretrained and configured VGG-19 model. A sample video with a frame of 128 × 128 was used to compute the computational time. For both models, video acquisition takes an average of 3.39 seconds to capture a single frame, and single- face detection and resizing takes 8.78 and 0.10 seconds, respectively. The configured model was approximately 1-second slower in emotion recognition, taking an average of 1.4 seconds to label the emotion of a single facial image; this drop in speed is relatively trivial, but likely reflects the increased processing demand associated with the higher-order parameters in the configured model. In total, the proposed framework takes 13.72 seconds to process and label the emotion of a single face within one frame. This computational speed may be sufficient for online feedback of emotions during class, which may assist teachers in making in-class decisions to manage and improve students' learning experiences. The proposed model may also be utilized in identifying the emotional climate in classrooms by aggregating classification outcomes across students. However, it is likely that this group-level computation will take additional time, and researchers interested in developing continuous or online measurements of classroom emotions will need to consider how to increase the speed and efficiency of this model further. The present results contribute to the that by showing that transfer learning can greatly increase CNN model performance with a negligible impact on processing speed.

Module	Pre-trained VGG-19	Configured VGG-19
Video frame acquisition	3.39	3.39
Face detection	8.78	8.78
Face resize	0.10	0.10
Emotion recognition	0.52	1.44
Total	**12.80**	**13.72**

Table 4: Computational time (in seconds) of each block averaged over 100 frames with a single face.

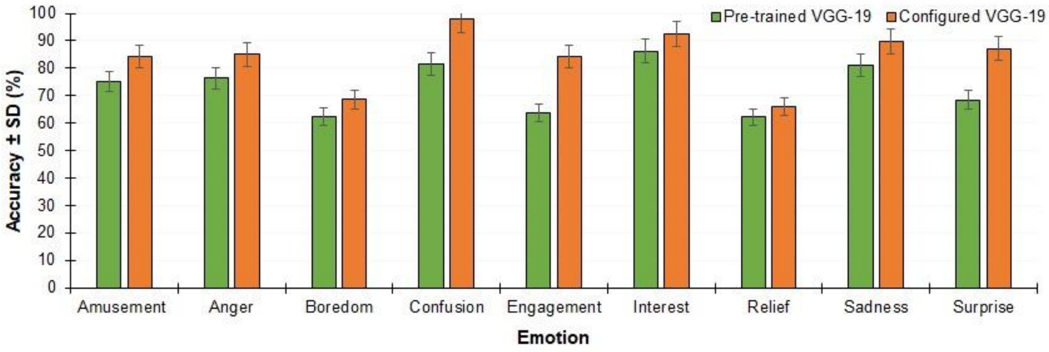

Figure 3: Comparison of average classification accuracy ± SD (%) performance in each facial expression category.

Conclusion

In this paper, a classroom emotion classification method based on deep transfer CNN framework was proposed and tested on a broad image dataset involving facial images of students in real classroom learning environments. The framework consisted of a pre-trained

VGG-19 and a configured VGG-19 CNN, where the pre-trained CNN is trained using natural images and the configured CNN is fine-tuned using facial expression images. The performance of the framework is evaluated on a constructed dataset integrating three publicly available and coded datasets that link academic emotions to students' facial expressions (i.e., DAiSEE, Raf-DB, and EmotioNet), as well as classroom videos from the internet. From the experimental results, it can be concluded that the con- figured VGG-19 CNN achieved better classification accuracy and sensitivity compared to pre-trained VGG-19. Configured VGG-19 performance was also better for detecting each emotion and these model improvements were achieved with negligible changes in classification speed. These outcomes demonstrate the viability of using transfer learning to create powerful machine learning models for specific classification tasks that may lack the required training data. The specific model proposed in this research may also help teachers to monitor the emotional climate in the classroom and make necessary adjustments to the lesson delivery for improving the engagement and learning outcomes. Future work includes improve system performance by increasing the dataset size, add other emotion categories or even compound emotion categories and improve robustness of the proposed methodology by combining student's behavioral cues (e.g., head pose) with academic affective states.

References

Ashwin, T., & Guddeti, R. M. R. (2020). Affective database for e-learning and classroom environments using indian students' faces, hand gestures and body postures. *Future Generation Computer Systems*, *108*, 334– 348.

Bajpai, R., Yuvaraj, R., & Prince, A. A. (2021). Automated eeg pathology detection based on different convolutional neural network models: Deep learning approach. *Computers in Biology and Medicine*, *133*, 104434.

Fabian Benitez-Quiroz, C., Srinivasan, R., & Martinez, A. M. (2016). Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 5562–5570).

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, *187*, 27–48.

Gupta, A., D'Cunha, A., Awasthi, K., & Balasubramanian, V. (2016). Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885*.

Hu, Q., Mei, C., Jiang, F., Shen, R., Zhang, Y., Wang, C., & Zhang, J. (2020). Rfau: A database for facial action unit analysis in real classrooms. *IEEE Transactions on Affective Computing*.

Li, S., Deng, W., & Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2852–2861).

Oh, S. L., Hagiwara, Y., Raghavendra, U., Yuvaraj, R., Arunkumar, N., Murugappan, M., & Acharya, U. R. (2020). A deep learning approach for parkinson's disease diagnosis from eeg signals. *Neural Computing and Applications*, *32*(15), 10927–10933.

Pabba, C., & Kumar, P. (2022). An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. *Expert Systems*, *39*(1), e12839.

Pekrun, R. (2014). Emotions and learning. educational practices series-24. *UNESCO International Bureau of Education*.

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The achievement emotions questionnaire (aeq). *Contemporary educational psychology*, *36*(1), 36–48.

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational psychologist*, *37*(2), 91–105.

Raghu, S., Sriraam, N., Temel, Y., Rao, S. V., & Kubben, P. L. (2020). Eeg based multi-class seizure type classication using convolutional neural network and transfer learning. *Neural Networks*, *124*, 202–212.

Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of educational psychology*, *104*(3), 700.

Skinner, E., Pitzer, J., & Brule, H. (2014). The role of emotion in engagement, coping, and the development of motivational resilience. *International handbook of emotions in education*, 331–347.

Thomas, J., Jin, J., Thangavel, P., Bagheri, E., Yuvaraj, R., Dauwels, J., . . . Westover, B. (2020). Automated detection of interictal epileptiform discharges from scalp electroencephalograms by convolutional neural networks. *International journal of neural systems*, *30*(11), 2050030.

Xu, G., Shen, X., Chen, S., Zong, Y., Zhang, C., Yue, H., . . . Che, W. (2019). A deep transfer convolutional neural network framework for eeg signal classication. *IEEE Access*, *7*, 112767–112776.

Yadegaridehkordi, E., Noor, N. F. B. M., Ayub, M. N. B., Affal, H. B., & Hussin, N. B. (2019). Affective computing in education: A systematic review and future research. *Computers & Education*, *142*, 103649.

Yuan, Q. (2022). Research on classroom emotion recognition algorithm based on visual emotion classication. *Computational Intelligence and Neuroscience*, *2022*.

Zeng, H., Shu, X., Wang, Y., Wang, Y., Zhang, L., Pong, T.-C., & Qu, H. (2020). Emotioncues: Emotion-oriented visual summarization of classroom videos. *IEEE transactions on visualization and computer graphics*, *27*(7), 3168–3181.

Contact email: yuvaraj.rajamanickam@nie.edu.sg