Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 9 No. 3 (2025) 551 - 561 e-ISSN: 2580-0760

Improving Frame-based Engagement Classification in E-Learning Using EfficientNet and Normalized Loss Weighting

Joseph A. Sugihdharma[1], Fitra Abdurrachman Bachtiar[2]*, Novanto Yudistira[3]
[1,2,3]Intelligent System Laboratory, Faculty of Computer Science, Brawijaya University, Malang, Indonesia

[1]josephananda88@student.ub.ac.id, [2]fitra.bachtiar@ub.ac.id, [3]yudistira@ub.ac.id

Abstract

Engagement can be defined as how individuals are involved in and interact with a task that requires attention and emotional conditions. Engagement is an affective state positively correlated with learning processes. Engagement along with other affective states, such as boredom, confusion, and frustration must be analyzed to identify students' learning behavior. Implementing proper prevention by measuring student engagement levels could increase students' learning intake. Such implementation involves building an effective feedback system or rearranging the learning design. Several researchers have proposed deep-learning approaches using the DAiSEE dataset to classify student engagement levels. In addition, previous studies utilized various loss functions equipped with class weighting to assign higher importance to the minor classes, which are low and very low engagement classes. Most of the state-of-the-art models achieved high accuracy, but the f1-score was still low because of the minor class struggle. This research tries to solve engagement level classification on imbalance conditions by proposing a normalized loss function weighting based on the Inverse Class Frequency formula based on each class' instances to give more importance and focus to the classes and trained on Vanilla EfficientNet model rather than experimenting on more advanced model to keep the efficient and suit the memory constraint on the e-learning implementation. Based on the conducted experiments, the normalized ICF obtained the highest accuracy of 51.64% and weighted f1-score of 50.86%, which is superior to the standard ICF performance, which received 50.32% accuracy and weighted f1-score of 50.49% using the same settings.

Keywords: classification; deep learning; engagement; EfficientNet; normalized loss

How to Cite: J. A. Sugihdharma, F. Bachtiar, and N. Yudistira, "Improving Frame-based Engagement Classification in E-Learning Using EfficientNet and Normalized Loss Weighting", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 9, no. 3, pp. 551 - 561, Jun. 2025.
Permalink/DOI: https://doi.org/10.29207/resti.v9i3.6161

Received: November 2, 2024
Accepted: June 8, 2025
Available Online: June 21, 2025

1. Introduction

Emotion is one of the important aspects in human social activities. Emotion combines various feeling, thoughts, and also human behaviour. Identification and emotion categorization which present and felt by someone can be done through emotion recognition. Various domains and applications have used emotion recognition as part of its process, such as marketing, e-learning, e-health, user experience, and even transportation and military [1]. In recent years, one of emotion recognition domain, namely e-learning, has grown rapidly through the presence of Massive Open Online Course which also known as MOOC. The growth and interest in this area has drawn researchers' attention to conduct research in this area.

Although there has been several research on e-learning topics, such as e-learning recommendation systems in [2] and e-learning personalization based on the user's learning style in [3], there are still few studies that discussing or classifying the student engagement in e-learning environment. As in [4], the term 'engagement' can be defined as how individuals are involved and interact with a task that requires attention and emotional state.

Researchers in [5] mentioned that there are three dimensions or types of student engagement, namely emotional, cognitive, and behavioral engagement. Emotional engagement refers to the affective states of the students during an activity conducted in the classroom. Cognitive engagement measures students' motivation, effort, and strategy when faced with a problem or even failure. Behavioral engagement focuses on activities, both academic and social, that are considered crucial for achieving success in school and preventing dropout.

The rapid growth of MOOC is followed by the increasing number of new students registering and participating in the MOOC. In other side, most of the MOOCs have low retention rates followed by high level of dropouts, reaching 91-93% on the first assignment with the completion rates of 45% on the first assignment [6]. The course completion rates itself reaching 40% on several courses, but most of them only have less than 10% of completion rates [7]. Students think that there are two common drawbacks found in MOOC, which are 'lecture fatigue' because most of the course content dominated by videos and 'poor course design' because the course is lack of feedback or two-way interactivity between the teachers and students. Detecting engagement emotion is one of prevention step to tackle several academic problems, such as low academic performance, feeling of being isolated, and high dropout as stated in [8].

Two of the most common approach in engagement emotion classification are frame-based and sequence-based or video-level approach. Various models were proposed in the state-of-the-art research to solve students' engagement classification in e-learning environments, by using Convolutional Neural Networks-based model to capture spatial representations better or Temporal and Recurrent Neural Network-based model to consider the relationship between each frames better. In order to enhance the performance of those models, several researchers tried out different loss weighting function to classify better and give a higher attention to the minor classes. As in [9], the researchers proposed 3D DenseAttNet to solve engagement classification problem using sequence of frames. The researchers utilized several loss functions, such as Cross Entropy loss (CE) and Class Balanced (CB)-based loss function, such as CB-Cross Entropy loss (CB-CE), and CB-Focal Loss (CB-FL). Another research by [10] proposed a Vision Transformer model to solve engagement classification problem in single frame scenario. The researchers also used the same loss function as [9], such as CE, CB-CE, and CB-FL. Additional loss function, such as Focal Loss (FL), Cross Entropy (CE) with sample weights, and Focal Loss (FL) with sample weights were also utilized.

Based on the confusion matrix results displayed in [9] and [10], it can be concluded that each loss function has its own characteristics. The suitability of a loss function weighting depends on the use case of the weighting formula. CB loss, proposed by [11] designed to tackle imbalanced data classification problem by re-weighting the samples using inverse effective number of samples, which trying to ensure that there is no information overlap among data as the samples increases. As a result, overfitting risk could be reduced and also minority cases could be detected and classified correctly. However, in this situation, CB-based loss as tested in [9] and [10], only classify very low to none amount of data (mostly less than 10 instances) to the minority classes, which is low engagement. Additionally, based on the research conducted by [10], the model that enhanced with CB loss weighting which achieved the highest performance using patch of 32 tends to classify most of the data to one of the major classes. This causes the increasing amount of false positive on other classes. On the other hand, when Categorical Cross Entropy (CCE) enhanced with Inverse Class Frequency (ICF) sample weight, the model obtained lower accuracy but the correctly classified data distribution is more balanced using the patch of 32. In addition, more minority class samples are detected which achieved the highest accuracy rather than CB loss. Even though the minority classes samples that classified correctly is very low, it can be concluded that ICF loss weighting shows a promising opportunity to classify the instances better.

Based on the identified characteristics of each loss function weighting in previous research, a new loss function weighting is proposed by modifying the ICF weighting, named Normalized-ICF weighting. The proposed Normalized-ICF weighting normalizes the weight of each class by the highest weight from the class list. The proposed Normalized-ICF loss function weighting aims to improve the model performance on classifying the minor classes, which is class 0 (very low engagement) and class 1 (low engagement). Several tests and experiments also conducted to test the performance of the model on different scenarios, namely sampling interval test which test the model using different number of frames taken from each video and image augmentation test which test the model on different input image transformation.

Detecting engagement emotions could tackle several academic problems, such as high dropout. Previous research in this field has proposed various models and enhanced them with various loss function weighting, but most have low performance on minor class, such as in [10]. Detecting the minor classes, namely very low and low engagement classes is important since the model will only classify the input image into high or very high engagement if the model is unable to detect the minor classes at all. This research aims to improve the performance of minor class detection by proposing a new loss function weighting by normalizing the class weights based on the ICF formula using a vanilla EfficientNet-B0 model. This model was chosen since it's one of the smallest and most efficient models in CNN-based architecture, which is suitable to be implemented in real-world devices. Rather than experimenting with models' architecture, one of the main focuses in this research is to find and tune the supporting components of the model that affect the model's perspective and performance during the training flow while keeping the architecture efficient for future implementation. Subsequently, several tests such as the sampling interval test and image augmentation test were also conducted. Sampling interval tests were intended to find the optimal interval that may achieve the best classification performance and minimize the redundancy in data sampling since the DAiSEE dataset

has very low movement on each frame and there are no certain rules that state the most optimal interval, while the image augmentations tests were conducted to find which augmentation could bring optimal variations to the model to maximize the classification performance.

2. Methods

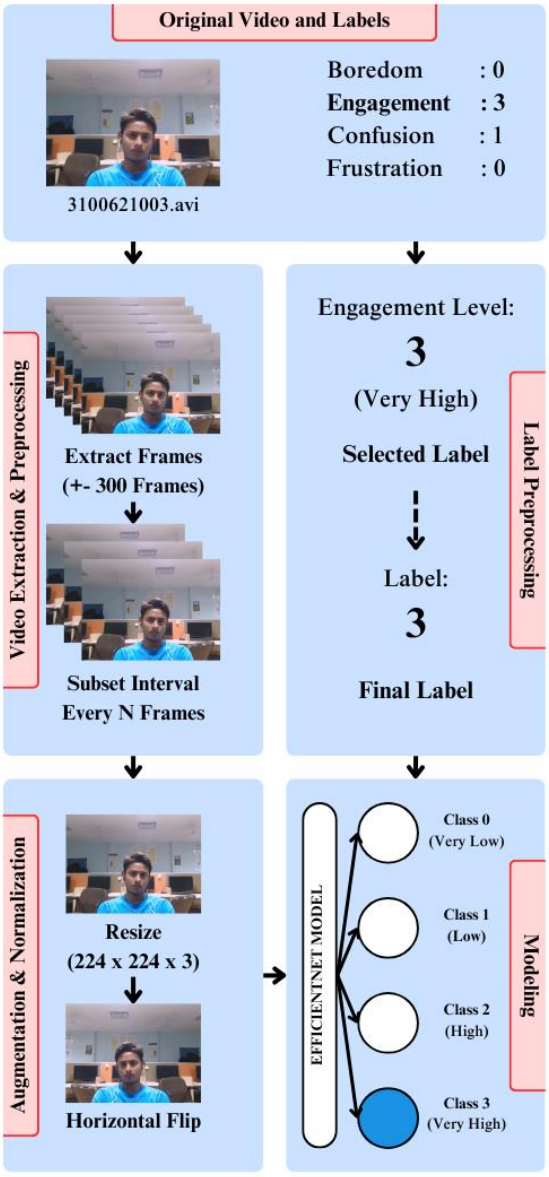


Figure 1. Methodology

Figure 1 shows the proposed methodology for engagement classification in this research. First, each video in DAiSEE dataset were extracted into single frames. Each video yield around 300 frames since it has a duration of 10 seconds each and 30 frame per second (fps). After each video were extracted, only several frames were taken to represent each video based on the defined frame-taking interval. Selected frames were merged into one big dataset and then shuffled since it is a frame-level engagement classification. On each frame, augmentations were done to increase variations on the dataset and each frame were resized to suitable size for the model. Furthermore, the data were normalized with ImageNet mean and standard deviation for the EfficientNet model and the class weights were counted based on the amount of data in each class. The class weighs will be used as the loss function weighting to give higher attention to the class with less data or higher weight. For the label, only engagement level will be considered and selected as the final label. After the frame and label are preprocessed, both will be passed into the chosen model to be trained. Finally, the performance of the trained model on each test were collected and analyzed to see the pattern and characteristics of the proposed loss function weighting. The tests are consists of several intervals test and image augmentations. The model is evaluated based on the accuracy and f1-score from the confusion matrix value using the test split provided by the dataset.

2.1 Dataset

DAiSEE (Dataset for Affective States in E-Environments) dataset is used in this study. This dataset was proposed by [12] from Indian Institute of Technology Hyderabad (IITH) and consists of 9.068 videos collected from 112 subjects ranging between 18 to 30 years old. Each video has a duration of 10 seconds with 30 frame per second (fps), which produce around 300 frames in total. The researchers in [12] identified several problems in previously available datasets, such as no datasets addresses four affective states which are necessary in e-learning environment and also most of them were taken in a controlled environment. Therefore, DAiSEE tried to tackled those problem by take the video in an uncontrolled environment or also called 'in-the-wild' and labelling each video with four affective states, namely boredom, engagement, confusion, and frustration. Each affective states have four intensity levels, ranging from 0 to 3 representing very low, low, high, and very high.

Table 1. DAiSEE class distribution

Class	Train Split	Val. Split	Test Split
Class 0 (Very Low)	34	23	4
Class 1 (Low)	213	143	84
Class 2 (High)	2617	813	882
Class 3 (Very High)	2494	450	814

Table 1 shows the distribution of each engagement level in the DAiSEE dataset on each split. Even though class 0 has less than 100 instances in total, but this class is retained based on their defined characteristics and in order to be fairly compared with other state-of-the-art researchs. According to [13] and [14], each engagement level can be distinguished based on several characteristics. Students with very low engagement (class 0) tends to tilt their body to the back, the eye gaze is not focused on the monitor [14], sometimes the eyes are completely closed, and not thinking about the given task [13]. Students with low engagement (class 1) barely opened their eyes, still not 'into' their task [13], and sometimes touched their eyelids unnecessarily, looking tired and frustrated [14]. Students with high engagement (class 2) show a calm expression, eyes focused on the monitor, and sit straightly [14]. They

don't need to be reminded to focus on the task [13]. Students with very high engagement (class 3) should be praised for their engagement in the task [13]. They show more seriousness and focused eyes on the screen [14].

2.2 Frame interval selections

There has been only few research that conduct engagement classification using frame-based approach. Just like in sequence-based approach, there has been no certain strict rules on how many frames taken for training or testing the classifier. Gupta et al. [12] as the baseline research used interval of 4 frames. In the other side, Mandia et al. [10] used interval of 30 frames, taking 10 frames to represent each videos. In the sequence-based or video-level approach, the researchers in [9] sampled the video using interval of 10 frames to classify the engagement level, while in [15] the researchers experimenting with 20, 40, 60, and 80 frames for each video snippet. In the baseline research, Gupta et al. [12] stated that there's no effect on accuracy when selecting different interval during the data sampling process. In order to prove this hypothesis and giving quantitative insights, 15 different intervals were tested in this research using systematic sampling strategy as shown in Equation 1.

$$k = \frac{N}{n} \quad (1)$$

The k variable represent the sample that will be selected from each video, while N represents the total frame in the video, and n is the sampling interval which will take a frame every n-th frame. The intervals or n were selected based on the multiplicative factor of 300 frames, which is the total amount of frames in a video. The selected intervals are sampled every 1, 2, 3, 4, 5, 6, 10, 15, 20, 30, 50, 60, 100, 150, and 300 frames from each video. Based on the selected interval value, number of samples k representing each video clip would be 300, 150, 100, 75, 60, 50, 30, 20, 10, 6, 5, 3, 2, and 1 respectively. The different sampling interval were done in the training set, while the test set used 30 frames sampling interval to match the number of the instances in Mandia et al. [10] for a fair comparison. Using all the frames in the dataset or sampling the frames consecutively may bring data redundancy since there are very little movement in each DAiSEE dataset video. Therefore, to introduce more generalization to the model, the frames were selected every n-th frame. Besides, sampling the data can reduce the computational cost rather than using all of the data.

2.3 Image augmentation combinations

Image augmentation tests were conducted in order to understand the model behavior and performance using different augmentation combinations. According to [16], image augmentations can be considered as data-level approach to tackle imbalance problem in the dataset. Two main types in data augmentations are data warping which transforms the image 'on the fly' and oversampling which create synthetic data based on the available dataset. Oversampling can increase the final dataset depending on how many factors that the data is oversampled. Based on the research that has been done by [17], it can be concluded that augmentation in data-space such as data warping transformation produces better performance as long as the chosen transformation is able to maintain the original label. Therefore, data warping approach were used in this study. During the data warping process, each image was transformed and augmented 'on the fly' during the training process, resulting in variations of the data to be seen by the model at each epoch during the training phase, but the actual stored dataset size and instances remain the same. As in [16], there are various augmentations types that can be chosen to tackle this problem from basic image augmentations up to deep learning approaches. Deep learning approaches covers adversarial training, neural style transfer, and Generative Adversarial Networks (GAN) which can generate new synthetic examples, but this approach is cost expensive and requires high computational power [18]. On the other side, basic image manipulations, such as geometric transformation, color transformations, random erasing, kernel filters, and mixing image are easier to implement. However, its suitability is domain-dependent and need a careful consideration regarding the 'safety' of the augmentation.

In the engagement classification context, not all of the mentioned approach might improve the classification performance, for example mixing images might make the final output have a little sense from human perspectives, random erasing potentially erases the important part of the picture such as faces, while color transformation might change the brightness of the environment even though the original pictures already captured in dark environments. Therefore, basic image transformations such as scaling, cropping, flipping, rotation, and shearing might become the most suitable image augmentation approach. In this research, only scaling, cropping, and flipping which are selected as image augmentation transformations, since rotation and shearing might make biases to the model during the prediction phase since the body posture affect the engagement level according to [14].

Based on the selected image augmentation transformations, there are five combinations used in this research to create variations of the data to the model during training, namely resize (scaling), random resized crop (scaling and cropping), resize + horizontal flip (scaling and flipping), random resized crop + horizontal flip (scaling, cropping, and flipping), and horizontal flip + random resized crop (scaling, cropping, and flipping). Two operations look identic to each other, namely 'random resized crop + horizontal flip' and 'horizontal flip + random resized crop', but two of them are different in implementation. They differ in the order in which the transformations are applied. In the random resized crop + horizontal flip operation, the image is randomly cropped, resized to a suitable input size for the model, and then flipped. In contrast, the horizontal flip + random resized crop transformations will flip the

image first, then crop a random area on the image, and finally resize the image.

2.4 State-of-the-art loss function weighting

There has been various approach in engagement emotion classification. Two of the most common approach in engagement emotion classification are frame-based and sequence-based approach. Currently, there are two state-of-the-art research besides the baseline research by Gupta et al. [12] that classified the engagement emotion using a frame-based approach, namely Mandia, Singh, and Mitharwal research in [9] as well as Adyapady and Annappa research [19]. On the other hand, there have been several research that classifying the engagement level using temporal approach, such as 3D based approach as in [9], Long Recurrent Convolutional Network (LRCN) as in [12], enhancing CNN based model with LSTM or TCN as in [20], or even passing images in the sequence one by one as in [15]. Based on the state-of-the-art research, there are two commonly used loss function weighting, namely ICF and CB.

ICF is one of the most common loss function weighting used in several state-of-the-art research. The ICF loss prevents the minority class samples to be classified into the majority classes [20] due to the imbalanced data distribution by giving a higher importance to the minority classes samples by assigning higher weights.

$$\alpha_i = \frac{1}{K_o} * \frac{K}{O} \tag{2}$$

As in [10], the standard ICF weighting formula is shown in Equation 2, where α_o represents the weighting in a loss function for class o, K_o represents the number of samples in class o, K represents total number of samples, and O represents total number of samples. The ICF weight is determined based on the proportions of samples in a certain class to all the samples across all classes. The weights of ICF commonly multiplied with the CCE loss.

As in [10], researchers utilized standard ICF as one of the loss function weightings for the proposed Vision Transformer model to classify student engagement level. The proposed Vision Transformer model achieved 49.86% accuracy when CCE with ICF sample weights is used and 54.87% accuracy without using sample weights using patch of 32, while lower accuracy of 52.16% and 46.49% obtained using CCE without sample weight and CCE with ICF sample weight respectively using patch of 64. Based on the provided confusion matrix, CCE without sample weight failed to classify any of the minority classes samples from class 0 or class 1 using the patch of 32, but the CCE equipped with ICF sample weights could classify 13 samples that belongs to class 1.

Class Balanced (CB) loss is also one of the loss function weightings commonly found in the state-of-the-art research on engagement level classification. CB loss was proposed by [11] and designed to tackle the imbalanced data problem by assigning a weighting factor that inversely proportional to the effective number of samples. Effective number of samples can be defined as volumes of samples which approximately enough to train the model and minimize the data usage with similar or overlapping features at a time.

$$E_n = \frac{1-\beta^n}{1-\beta}, where\ \beta = \frac{(N-1)}{N} \tag{3}$$

$$CB(p,y) = \frac{1}{E_{n_y}} L(p,y) = \frac{1-\beta}{1-\beta^{n_y}} L(p,y) \tag{4}$$

The effective number of samples formula is shown in Equation 3. The n variable represents the amount of samples and β is a hyperparameter which has element value between 0 and 1 ($\beta \in [0,1]$). The CB loss formula was formed by performing inverse on the effective number of samples as shown in Equation 4, where $L(p,y)$ is an independent loss function which can be added to the CB loss function, while p is the probability of predicted class and y is the ground truth class label from the predicted sample. n_y represents the amount of samples n which has label of y.

$$CB_{softmax-CCE} = -\frac{1-\beta}{1-\beta^{n_y}} \log\left(\frac{e^{z_y}}{\sum_{j=1}^{C} e^{z_j}}\right) \tag{5}$$

$$CB_{FL} = -\frac{1-\beta}{1-\beta^{n_y}} \sum_{i=1}^{C} (1-p_i^t)^{\gamma} \log\left(p_i^t\right) \tag{6}$$

CB Loss commonly paired with loss function such as CCE to form CB-CCE as shown in Equation 5 or FL to form CB-FL as shown in Equation 6. In CB-CCE formula, the class balanced weighting is multiplied by the value of softmax CCE formula, while in the CB-FL formula the weighting is multiplied with the modulating factor of $(1-p_i^t)^{\gamma}$ and the sigmoid CCE formula. The p_i^t variable represent the logits from the sigmoid activation function.

As in [10], researchers also utilized CB loss weighting besides standard ICF for the proposed Vision Transformer model to classify student engagement level. The proposed Vision Transformer model achieved highest accuracy of 55.18% in patch of 32 using CB-CCE loss and 52.44% when CB-FL is used. Using patch of 64, the accuracy dropped to 51.32% when using CB-CCE and 45.95% using CB-FL. However, when CB is used, the accuracy is 3%-6% higher using patch of 32 compared to weighted CCE using ICF loss weighting, but the minority class detection in CB-FL and CB-CCE is lower than ICF, achieving true positive value of 6 and 0 in class 1 respectively.

2.5 Normalized loss function weighting

In this research, the proposed loss function weighting modifies the weighting formula used by [10], which is standard ICF formula. Weight normalization was performed in order to help the model become more sensitive to classify the minor classes and also put the weighting range between 0 and 1. Normalizing the value also helps to speed up the learning phase of the model [21]. The proposed formula for loss function weighting normalization is shown in Equation 7. In this

formula, the weight of the loss function which originally calculated with standard ICF formula will divided by the maximum number of the calculated weights between all classes. Therefore, the resulting weights will have value of 1 for the class with the highest weights, and anywhere between 0 and 1 for the rest of the class which originally have lower weights.

$$\alpha_i = \frac{\frac{1}{K_O} * \frac{K}{O}}{max\left(\frac{1}{K_O} * \frac{K}{O}\right)} \quad (7)$$

There have been several methods on how to normalize the value of a data list, such as normalizing a feature value using Min-Max normalization as shown in Equation 8. As in [21], the researchers compared several normalization method, namely Zero-Mean normalization, Sigmoidal normalization, Softmax normalization, and Min-Max normalization. The results obtained in [21] showed that Min-Max normalization has better performance and calculation time compared to other normalization method.

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (8)$$

Min-Max normalization works by substracting the current data X with the minimum value in the list then divided by the substraction between maximum value and minimum value in the list. The proposed normalized ICF weighting modifies the original Min-Max normalization. The approach of substracting current value and maximum value with the minimum value was not adopted in the proposed loss function shown in Equation 7. The removal of minimum value substraction prevent the class weight which has minimum value to be normalized to zero. When a class weight has value of zero, it might affect the whole loss calculation process and give zero importance on the corresponding class. Using the proposed normalized ICF formula, the weights of each class is normalized as shown in Table 2.

Table 2. Standard and normalized ICF weight

Class	Standard ICF Weight	Proposed Normalized ICF Weight
Class 0 (Very Low)	39.3485	1.0000
Class 1 (Low)	6.3076	0.1603
Class 2 (High)	0.5117	0.0130
Class 3 (Very High)	0.5371	0.0136

2.6 Evaluation metrics

In order to evaluate the classification performance from each model, several metrics were calculated based on the confusion matrix value. True positive (TP), false positive (FP), false negative (FN), and true negative (TN) were used to calculate accuracy, precision, recall, and f1-score.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (9)$$

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

$$F1 - Score = 2 \times \frac{precision \times recall}{precision + recall} \quad (12)$$

Accuracy can be calculated by dividing all true predicted instances with all of the data using the formula shown in Equation 9 to find out how many samples were correctly classified. Equation 10 shows the formula of precision which used to calculate how many positive samples were correctly classified among all of the samples that were predicted as positive. Recall can be calculated using the formula shown in Equation 11 to count how many positive samples were correctly classified among all the samples that belong to the corresponding class. f1-score can be measured using Equation 12, which calculates the harmonic mean of precision and recall. Based on the known evaluation metrics value, namelyprecision, recall, and f1-score, macro averaging and weighted averaging were calculated for the final value which represent the performance of all-class classification.

$$Macro\ Precision = \frac{1}{C}\sum_{i=1}^{C} Precision_i \quad (13)$$

$$Macro\ Recall = \frac{1}{C}\sum_{i=1}^{C} Recall_i \quad (14)$$

$$Macro\ F1 - Score = \frac{1}{C}\sum_{i=1}^{C} F1_i \quad (15)$$

$$Weighted\ Precision = \sum_{i=1}^{C} w_i \times Precision_i \quad (16)$$

$$Weighted\ Recall = \sum_{i=1}^{C} w_i \times Recall_i \quad (17)$$

$$Weighted\ F1 - Score = \sum_{i=1}^{C} w_i \times F1_i \quad (18)$$

The formulas in Equations 13 through 15 are used to calculate macro precision, macro recall, and macro f1-score respectively. The C variable represents the total class in the dataset. The macro-averaging sums up each value in the calculated evaluation metrics value from each class, then divides it by the total number of classes. The macro-averaging treats all classes equally, regardless of class frequency. The formulas in Equations 16 through 18 are used to calculate weighted precision, weighted recall, and weighted f1-score respectively. The w_i variable represents the proportion of the class frequency or true samples in each class divided by all of the samples in the dataset. In contrast to macro-averaging, weighted-averaging calculates the final value based on the proportion of each class frequency, rather than averaging them equally.

Besides macro-averaging, micro-averaging computes a global average f1-score by calculating the sum or total of the true positive (TP), false negative (FN), and false positive (FP) values across all classes. In other words, micro-averaging essentially computes the proportion of correctly classified samples out of all samples and this condition has the same behavior as calculating overall accuracy. In a case where the final classification is a single label, the micro precision, micro recall, and micro f1-score will have the same value with accuracy. Therefore, a single accuracy value is sufficient, since it is equivalent and could represent micro precision, micro recall, and micro f1-score in a single-label classification.

3. Results and Discussions

In this research, the performance of two loss function weighting were compared, namely state-of-the-art Mandia loss function weighting as defined in [9] and the proposed normalized loss function weighting. Several tests were carried out in different scenarios as defined in the previous section in order to understand and gain optimal performance for both of the loss function weighting, which are interval tests and augmentation test. Each test for both loss function weighting was done using default or vanilla EfficientNet-B0 model which is the smallest variant of EfficientNet model proposed by [22].

Table 3 shows the initial test result from both loss function weighting. Each loss function weighting was applied to EfficientNet model that was trained using frames interval of 30, resize + horizontal flip image augmentations, 10 epochs, batch size of 50, learning rate of 0.001 with gamma of 0.5 and step size of 5 which will cut the learning rate by 50% every 5 epochs. Based on the obtained results, it can be seen that our proposed normalized loss function weighting is superior in all evaluation metrics. It obtained accuracy of 51.64%, macro f1-score of 30.67%, and weighted f1-score of 50.86%.

Table 3. Initial test results

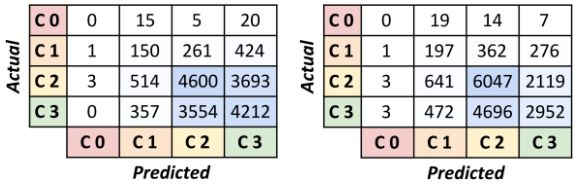
Intervals	Standard ICF	Normalized ICF (Proposed)
Accuracy	50.32%	51.64%
Macro Precision	29.89%	31.09%
Macro Recall	30.5%	32.14%
Macro F1-Score	30.14%	30.67%
Weighted Precision	30.72%	52.75%
Weighted Recall	50.33%	51.64%
Weighted F1-Score	50.49%	50.86%

Fifteen sampling intervals are defined and tested on both EfficientNet model with standard ICF loss weighting and proposed normalized ICF loss weighting. Different intervals were selected to understand the performance of each loss function weighting when trained with different amount of data. The model was trained using 10 epochs, batch size of 50, and Resize + Random Horizontal Flip augmentation. The accuracy and f1-score of the model on different interval scenario shown in Table 4.

Based on the obtained weighted f1-score, it can be seen that both model performs well when trained and tested using interval of 30, which consists of 10 frames that represent each video, achieving 50.49% using standard ICF loss function weighting and 50.86% using normalized ICF loss function weighting. From 15 tested interval, model with normalized ICF loss weighting superior in 9 intervals, while standard ICF only achieve higher accuracy in 6 intervals. On some intervals, there's a big difference in performance value between both of the loss function weightings. Based on the state-of-the-art research, the intervals used in [10] is interval of 30 and [12] sampled every 4th frame when utilizing EmotionNet to analyze DAiSEE. Therefore, a deeper look is done based on the confusion matrix of both intervals.

Table 4. Interval test accuracy

Intervals (frame)	Standard ICF		Normalized ICF	
	Acc	Weigh. F1	Acc	Weigh. F1
1	45.92%	45.79%	47.73%	46.88%
2	45.15%	44.28%	47.16%	46.42%
3	50.12%	49.1%	47.62%	46.35%
4	48.75%	45.92%	51.85%	50.27%
5	46.8%	45.68%	49.3%	48.54%
6	50.1%	47.62%	52.17%	49.75%
10	49.83%	48.56%	49.1%	46.98%
15	48.98%	47.49%	50.14%	49.2%
20	47.71%	47.22%	48.94%	48.26%
30	50.32%	50.49%	51.64%	50.86%
60	47.72%	47.3%	47.52%	47.29%
75	50.31%	50.11%	50.44%	49.51%
100	50.46%	50.09%	48.2%	48.01%
150	44.52%	43.91%	44.2%	43.62%
300	46.7%	44.83%	46.75%	45.08%



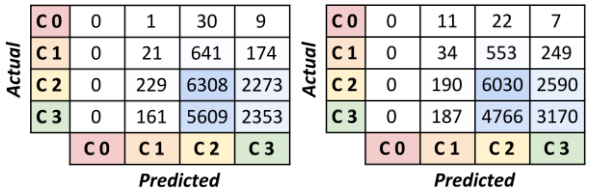
(a) (b)

Figure 2. Interval of 30's confusion matrix (a) using standard ICF and (b) using normalized ICF

Figure 2 shows the results of the interval of 30 on both loss weightings. It can be seen that normalized ICF achieve higher true positive on class 1, which correctly classified 197 samples on the minority class. When majority class samples taken into considerations, model with standard ICF achieve more balanced distribution between class 2 and 3, while normalized ICF suffering similar phenomenon as the highest performing loss weighting in [10] which tends to classify most of the majority samples into class 2.

Sampling every 4th frame results the confusion matrix as shown in Figure 3. Generally, the true positive of the minority class decreased significantly on both loss weightings. The true positive value of class 1 using the normalized ICF is higher than using standard ICF, namely 34 samples were correctly classified into class 1. Besides, the total of false positive on the class 1 is also lower on model with normalized ICF weighting. On the majority samples of class 2 and 3, the true positive distribution is also more balanced on the model with normalized ICF rather than standard ICF.

Generally, it can be concluded that the selection of intervals can affect the model's performance significantly, especially when detecting minority samples. Interval of 30 achieve better performance on both models since it takes 10 frames to represents each video, which a moderate number of frames to prevent frame similarity since there's very minimum movement on each frame. Since interval of 30 achieve the highest f1-score, this interval is brought into the next test, which is image augmentation test.



(a) (b)

Figure 3. Interval of 4's confusion matrix (a) using standard ICF and (b) using normalized ICF

Table 5. Image augmentation test accuracy

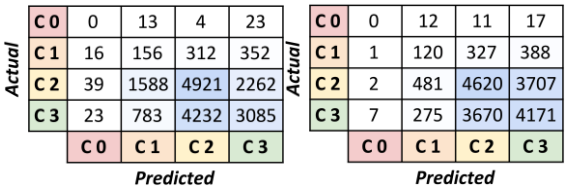
LR	Image Augmentation	Standard ICF	Normalized ICF (Proposed)
0.001	Resize	**52.01%**	46.15%
0.001	Random Resized Crop	**48.28%**	46.15%
0.001	Resize + Horizontal Flip	50.32%	**51.64%**
0.001	Random Resized Crop + Horizontal Flip	**44.71%**	42.17%
0.001	Horizontal Flip + Random Resized Crop	45.83%	**50.04%**
0.0001	Resize	45.24%	**45.25%**
0.0001	Random Resized Crop	**42.06%**	42.03%
0.0001	Resize + Horizontal Flip	46.55%	**46.58%**
0.0001	Random Resized Crop + Horizontal Flip	42.8%	42.8%
0.0001	Horizontal Flip + Random Resized Crop	**43.92%**	43.89%

Table 6 and Table 6 shows the accuracy and weighted f1-score of image augmentations test on both loss weightings respectively. There are five combinations of augmentations, each tested on two different learning rates, namely 0.001 and 0.0001. Generally, most of the defined augmentations yield similar results in terms of accuracy and f1-score when the models compared to each other, but not all of them obtained similar confusion matrix value. For example, the Resize + Horizontal Flip using 0.001 learning rate obtained different minority class true positive value on both models as shown in Figure 2. Therefore, a deeper look on the confusion matrix value is needed.

Confusion matrix in Figure 4 shows the results on both model when the data is transformed using horizontal flip and random resized crop. Based on the confusion matrix value, it can be seen that the true positive of class 1 using standard ICF is higher than normalized ICF, but the false positive of class 2 sample being classified as class 1 is reaching more than 1000 samples. On the majority class, it seems that higher number of samples is classified correctly using normalized ICF, resulting in higher accuracy of 50.04% with 50.06% f1-score. The imbalanced in majority class prediction using standard ICF is mainly caused by the high false negative rates of class 3 which is classified as class 2 samples.

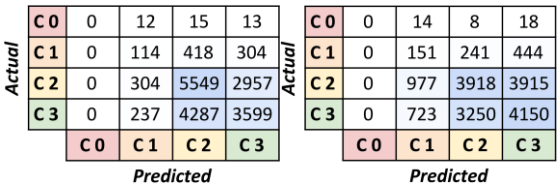
Table 6. Image augmentation test weighted f1-score

LR	Image Augmentation	Standard ICF	Normalized ICF (Proposed)
0.001	Resize	**51.38%**	47.15%
0.001	Random Resized Crop	**49.03%**	48.37%
0.001	Resize + Horizontal Flip	50.49%	**50.86%**
0.001	Random Resized Crop + Horizontal Flip	**44.82%**	42.99%
0.001	Horizontal Flip + Random Resized Crop	47.4%	**50.06%**
0.0001	Resize	44.71%	44.71%
0.0001	Random Resized Crop	**39.28%**	39.25%
0.0001	Resize + Horizontal Flip	46.39%	**46.42%**
0.0001	Random Resized Crop + Horizontal Flip	39.41%	**39.42%**
0.0001	Horizontal Flip + Random Resized Crop	**40.93%**	40.89%



(a) (b)

Figure 4. Horizontal flip + random resized crop augmentation with 0.001 learning rate (a) using standard ICF and (b) using normalized ICF



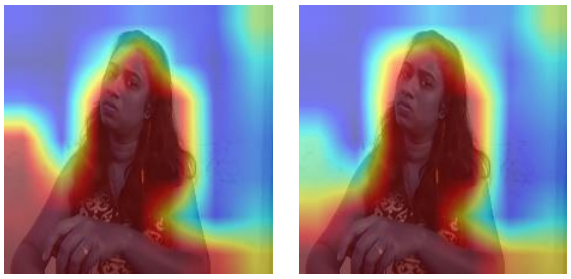
(a) (b)

Figure 5. Resize augmentation with 0.001 learning rate (a) using standard ICF and (b) using normalized ICF

Another confusion matrix as shown in Figure 5 is taken into deeper analysis. The confusion matrix in Figure 5 is the results obtained using resize augmentation and has opposite performance with the previous confusion matrix, where standard ICF shows higher performance rather than normalized ICF. From the minority class true positive, it can be seen that normalized ICF still dominating the results, classifying 151 class 1 samples correctly. Based on the majority class samples classification, it can be seen that normalized ICF prediction is lower than standard ICF, but obtained more balanced distribution. A cross-comparison can be done by comparing normalized ICF weighting when trained using resize augmentation with standard ICF weighting when trained using random resized crop + horizontal flip since it achieved similar performance on accuracy and also f1-score. Minority class which classified correctly seems similar between both model, but normalized ICF performed better classification

since the number of false positive on samples predicted as class 1 is much lower than standard ICF.

Additional observations can be seen on the confusion matrix of each model which obtained highest weighted f1-score using a certain image augmentation. Model with standard ICF achieved highest weighted f1-score of 51.38% using resize augmentation and 0.001 learning rate, resulting confusion matrix as shown in Figure 5(a). On the other hand, model with proposed normalized ICF achieved highest weighted f1-score of 50.86% using resize + horizontal flip and 0.001 learning rate, resulting confusion matrix as shown in Figure 2(b). While the weighted f1-score of normalized ICF model is a bit lower rather than standard ICF model, but the correctly classified instances of class 1 samples are higher, namely 197 instances compared with the standard ICF model which correctly classified only 114 class 1 instances. Based on the observations of the model through several testing, it can be concluded that various sampling interval and various image augmentations may perform differently on the model final performance. The model may perform well with a certain sampling and augmentation, but also may perform lower when different sampling or augmentations applied. While sampling interval and image augmentations techniques can help improving overall classification performance in highly imbalanced dataset such as DAiSEE, there are some trade-offs that have to be taken into consideration based on the main purpose of the research, such as the true positive in certain class, accuracy, and f1-score. In order to observe each model deeper, Gradient-weighted Class Activation Maps (GradCAM) is utilized to understand the region that gains the models' attention the most. Two images from different video clip randomly taken and shown in Figure 6 and Figure 7.

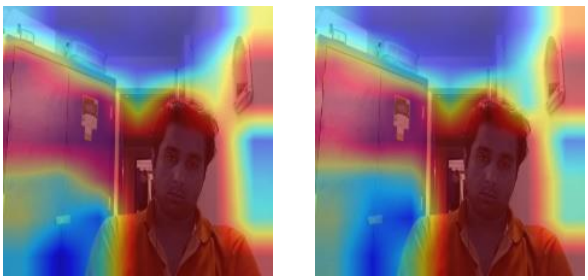


(a) (b)

Figure 6. GradCAM result of video 8264120127.avi at the 150th frame: (a) using standard ICF weighting and (b) using proposed normalized ICF weighting

Based on the GradCAM results as shown in Figure 6, it can be seen that the model with standard ICF marks unnecessary area on the bottom left corner which marked with red colour as an important part that determine the levels of engagement. Oppositely, model with normalized ICF marks those area as less important.

Another example of GradCAM evaluation on both loss function weighting shown in Figure 7. Although both of them looks similar, but it can be seen that standard ICF give attention to a wider area on the cupboard behind the subject, while proposed ICF give less or thinner attention on the cupboard and on the space between the mirror and the subject. Similar phenomenon can be seen on the blank space between the subject's head and the mirror behind the subject.



(a) (b)

Figure 7. GradCAM result of video 5100342042.avi at the 150th frame: (a) using standard ICF weighting and (b) using proposed normalized ICF weighting

Table 7. State-of-the-art (Frame Level)

Methods	Accuracy	Weighted F1-Score
InceptionV3 [12]	47.10%	Unknown
Facial Engagement Analysis Net (FEA-Net) [19]	**62.16%**	39.54%
Vision Transformer [10]	55.18%	49.41%
EfficientNet-B0 with normalized ICF loss weighting (proposed)	51.64%	**50.86%**

Table 7 shows the accuracy and f1-score between the proposed method, namely EfficientNet-B0 with normalized loss function weighting compared to the previous state-of-the-art level in frame-based engagement classification. The original research of the dataset written by [12], obtained a baseline accuracy of 47.10%, while the f1-score is unknown since the confusion matrix value wasn't provided by the researchers. As in [19], the researchers utilized two data types extracted from a single image, namely image and facial features extracted using MTCNN, such as facial landmarks, head pose estimation, eye gaze estimation, and facial action unit. An accuracy of 62.16% was obtained and it was the highest accuracy among other frame-level research, but only achieved a weighted f1-score value of 39.54%. A more advanced model utilized by [10] to solve this problem, namely the Vision Transformer. However, the model only achieved 55.18% accuracy and a weighted f1-score of 49.41%. The model still struggling to achieve better evaluation metrics value. The proposed model in this research utilized the smallest variant of EfficientNet, which is EfficientNet-B0. The proposed model achieved the highest accuracy of 50.49% and a weighted f1-score of 50.86%. Numerically, the performance of EfficientNet in this research slightly overpass the performance of Vision Transformer in [10] in terms of weighted f1-score metrics, which is 1.45% higher, but those numbers become more significant if the model efficiency is taken into consideration.

As shown in Table 8, the proposed EfficientNet-B0 becomes really efficient in terms of parameters amount, Giga Floating Operations (GFlops), and the model size compared to the previous models. The value of each

aspect was collected from PyTorch documentation of each model. The proposed model has 3 million fewer parameters and two times smaller GFlops than the InceptionV3 model but can achieve 7% higher accuracy. EfficientNet-B0 is also 20 times smaller than Vision Transformer, but able to achieve a 3% higher weighted f1-score. EfficientNet-B0 was chosen in this research since it's the smallest variant and one of the smallest architectures compared with other CNN models, such as InceptionV3 or ResNet. Rather than experimenting with models' architecture, one of the main focuses in this research is to find and tune the supporting components of the model that affect the model's perspective and performance during the training flow.

Table 8. State-of-the-art Model Efficiency (Frame Level)

Methods	Params	GFlops	Size
InceptionV3 [12]	27.1M	5.71	103.9MB
Vision Transformer [10]	85.6M	17.56	330.3MB
EfficientNet-B0 with normalized ICF loss weighting (proposed)	**24.1M**	**2.34**	**16MB**

Because of its efficiency and small model size, EfficientNet-B0 is more suitable to be embedded or integrated into the e-learning platform to provide real-time frame-based engagement level classification. If a resource constraint made the real-time classification heavy, the interval of students' face capturing can be adjusted, for example, the image of the students will be captured every 5 seconds or even a longer interval. Research with sequence-based or video-level approach such as [9], considers the decision of engagement levels based on several frames. This approach could be more accurate in terms of reducing the bias that occurred in single-frame cases, such as students classified with high engagement levels but accidentally closed their eyes or blinking in the observed timeframe. On the other hand, the sequence-based or video-level approach needs to stack several frames to the memory before a final decision is made, which means if there are resources or memory constraints, this approach is less suitable for this case. Therefore, further research at the frame level must be conducted since this approach is more suitable if the resources are limited in the end systems and also because this approach is less considered to be conducted since it's challenging and difficult to improve the model performance.

4. Conclusions

Inverse Class Frequency is one of the loss function weightings which commonly used in engagement level classification and performed pretty well on classifying samples especially in minority classes. In this research, a new normalized ICF is proposed to be more sensitive and performed better especially on minor classes. When applied to the real-world applications, this approach could prevent the model to classify the engagement only to class 2 (high engagement) and class 3 (very high engagement). Based on the conducted experiments, it can be seen that normalized ICF could reach the best performance using resize + horizontal flip augmentation, resulting 51.64% in accuracy and 50.86% in f1-score. Besides, normalized ICF also dominating in reaching higher f1-score on 9 different intervals, which is higher than standard ICF. It can be concluded that normalized ICF performed generally better rather than standard ICF in engagement level classification using DAiSEE dataset. The main focus and contribution of this research is finding how to keep the model small and efficient while also maintaining and achieving higher classification performance. This unique approach is used as a new perspective to explore the supporting component of the model which affects the training flow performance rather than experimenting on advanced and bigger model architecture.

In future research, more testing on the proposed normalized loss function weighting needs to be conducted to understand how robust or general the weighting is across various datasets and models. Modification of the formula or combination with other types of loss weighting such as Class Balanced (CB) also possible to be conducted in further research. After a sufficient and reasonable range of accuracy is achieved, implementation and integration in the end systems and devices can be performed to determine the model performance and behavior in real-environment testing.

Acknowledgements

The authors declare no conflict of interest. The author wishes to extend sincere appreciation to the Research and Service Community Institute (LPPM) at Brawijaya University for their invaluable support. This study received support from the Implementation of the State University Operational Assistance Program Research Program Number 00309.89/UN10.10501/B/PT.01.03/2/2024.

References

[1] J. Zhang, Z. Yin, P. Chen, dan S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Inf. Fusion*, vol. 59, hal. 103–126, Jul 2020, doi: 10.1016/j.inffus.2020.01.011.

[2] K. K. Jena *et al.*, "E-Learning Course Recommender System Using Collaborative Filtering Models," *Electronics*, vol. 12, no. 1, hal. 157, Des 2022, doi: 10.3390/electronics12010157.

[3] K. A. Laksitowening, A. P. Yanuarifiani, dan Y. F. A. Wibowo, "Enhancing e-learning system to support learning style based personalization," in *2016 2nd International Conference on Science in Information Technology (ICSITech)*, Balikpapan, Indonesia: IEEE, Okt 2016, hal. 329–333. doi: 10.1109/ICSITech.2016.7852657.

[4] A. N. R. Paidja dan F. A. Bachtiar, "Engagement Emotion Classification through Facial Landmark Using Convolutional Neural Network," in *2022 2nd International Conference on Information Technology and Education (ICIT&E)*, Malang, Indonesia: IEEE, Jan 2022, hal. 234–239. doi: 10.1109/ICITE54466.2022.9759546.

[5] J. J. Appleton, S. L. Christenson, dan M. J. Furlong, "Student engagement with school: Critical conceptual and methodological issues of the construct," *Psychol. Sch.*, vol. 45, no. 5, hal. 369–386, Mei 2008, doi: 10.1002/pits.20303.

[6] S. Tu, "ENGAGEMENT PREDICTION AND

VISUALIZATION IN ONLINELEARNING," in *Jiangsu Annual Conference on Automation (JACA 2020)*, Zhenjiang, China: Institution of Engineering and Technology, 2021, hal. 57–62. doi: 10.1049/icp.2021.1428.

[7] M. Hu, H. Li, W. Deng, dan H. Guan, "Student Engagement: One of the Necessary Conditions for Online Learning," in *2016 International Conference on Educational Innovation through Technology (EITT)*, Tainan, Taiwan: IEEE, Sep 2016, hal. 122–126. doi: 10.1109/EITT.2016.31.

[8] H. A. El-Sabagh, "Adaptive e-learning environment based on learning styles and its impact on development students' engagement," *Int. J. Educ. Technol. High. Educ.*, vol. 18, no. 1, 2021, doi: 10.1186/s41239-021-00289-4.

[9] N. K. Mehta, S. S. Prasad, S. Saurav, R. Saini, dan S. Singh, "Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement," *Appl. Intell.*, vol. 52, no. 12, hal. 13803–13823, 2022, doi: 10.1007/s10489-022-03200-4.

[10] S. Mandia, K. Singh, dan R. Mitharwal, "Vision Transformer for Automatic Student Engagement Estimation," in *2022 IEEE 5th International Conference on Image Processing Applications and Systems (IPAS)*, Genova, Italy: IEEE, Des 2022, hal. 1–6. doi: 10.1109/IPAS55744.2022.10052945.

[11] Y. Cui, M. Jia, T. Y. Lin, Y. Song, dan S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Long Beach, USA, 2019, hal. 9260–9269. doi: 10.1109/CVPR.2019.00949.

[12] A. Gupta, A. D'Cunha, K. Awasthi, dan V. Balasubramanian, "DAiSEE: Towards User Engagement Recognition in the Wild," vol. 14, no. 8, hal. 1–12, 2016, [Daring]. Tersedia pada: http://arxiv.org/abs/1609.01885

[13] J. Whitehill, Z. Serpell, Y. C. Lin, A. Foster, dan J. R. Movellan, "The faces of engagement: Automatic recognition of student engagement from facial expressions," *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, hal. 86–98, 2014, doi: 10.1109/TAFFC.2014.2316163.

[14] L. Geng, M. Xu, Z. Wei, dan X. Zhou, "Learning Deep Spatiotemporal Feature for Engagement Recognition of Online Courses," in *2019 IEEE Symposium Series on Computational Intelligence, SSCI 2019*, Xiamen, China, 2019, hal. 442–447. doi: 10.1109/SSCI44817.2019.9002713.

[15] T. Selim, I. Elkabani, dan M. A. Abdou, "Students Engagement Level Detection in Online e-Learning Using Hybrid EfficientNetB7 Together With TCN, LSTM, and Bi-LSTM," *IEEE Access*, vol. 10, hal. 99573–99583, 2022, doi: 10.1109/ACCESS.2022.3206779.

[16] C. Shorten dan T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0197-0.

[17] S. C. Wong, A. Gatt, V. Stamatescu, dan M. D. McDonnell, "Understanding Data Augmentation for Classification: When to Warp?," *2016 Int. Conf. Digit. Image Comput. Tech. Appl. DICTA 2016*, 2016, doi: 10.1109/DICTA.2016.7797091.

[18] J. Zhang dan C. Li, "Adversarial Examples: Opportunities and Challenges," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, no. 7, hal. 2578–2593, 2020, doi: 10.1109/TNNLS.2019.2933524.

[19] R. Rashmi Adyapady dan B. Annappa, "Learning Engagement Assessment in MOOC Scenario," in *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT 2022)*, Bangalore, India: IEEE, 2022. doi: 10.1109/CONECCT55679.2022.9865699.

[20] A. Abedi dan S. S. Khan, "Improving state-of-the-art in Detecting Student Engagement with Resnet and TCN Hybrid Network," in *2021 18th Conference on Robots and Vision (CRV)*, IEEE, 2021, hal. 151–157. doi: 10.1109/CRV52889.2021.00028.

[21] W. Li dan Z. Liu, "A method of SVM with normalization in intrusion detection," *Procedia Environ. Sci.*, vol. 11, no. PART A, hal. 256–262, 2011, doi: 10.1016/j.proenv.2011.12.040.

[22] M. Tan dan Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, hal. 10691–10700, 2019.