Convolutional Neural Network Model based Students' Engagement Detection in Imbalanced DAiSEE Dataset

Mayanda Mega Santoni[1], T. Basaruddin[2], Kasiyah Junus[3]
Doctoral Student, Faculty of Computer Science, University of Indonesia, Depok, Indonesia[1]
Faculty of Computer Science, University of Indonesia, Depok, Indonesia[2, 3]

***Abstract*—The COVID-19 pandemic has significantly changed learning processes. Learning, which had generally been carried out face-to-face, has now turned online. This learning strategy has both advantages and challenges. On the bright side, online learning is unbound by space and time, allowing it to take place anywhere and anytime. On the other side, it faces a common challenge in the lack of direct interaction between educators and students, making it difficult to assess students' engagement during an online learning process. Therefore, it is necessary to conduct research with the aim of automatically detecting students' engagement during online learning. The data used in this research were derived from the DAiSEE dataset (Dataset for Affective States in E-Environments), which comprises ten-second video recordings of students. This dataset classifies engagement levels into four categories: low, very low, high, and very high. However, the issue of imbalanced data found in the DAiSEE dataset has yet to be addressed in previous research. This data imbalance can cause errors in the classification model, resulting in overfitting and underfitting of the model. In this study, Convolutional Neural Network, a deep learning model, was utilized for feature extraction on the DAiSEE dataset. The OpenFace library was used to perform facial landmark detection, head pose estimation, facial expression unit recognition, and eye gaze estimation. The pre-processing stages included data selection, dimensional reduction, and normalization. The PCA and SVD techniques were used for dimensional reduction. The data were later oversampled using the SMOTE algorithm. The training and testing data were distributed at an 80:20 ratio. The results obtained from this experiment exceeded the benchmark evaluation values on the DAiSEE dataset, achieving the best accuracy of 77.97% using the SVD dimensional reduction technique.**

Keywords*—Convolutional neural networks; imbalanced data; deep learning; PCA; COVID-19; online learning; students' engagement; SVD; SMOTE

I. Introduction

During the COVID-19 pandemic, the education sector has been compelled to adopt online learning. The conventional classroom learning has transformed into online learning or "school from home." E-learning has become a standard solution for learning, and virtual conference technologies, such as Zoom, Google Meet, and others, have given online learning flexibility and accessibility from anywhere and at any time, suitable with the current digital era. However, despite the numerous advantages of online learning, one significant obstacle that needs to be addressed is the lack of direct interaction between teachers and students. During virtual conferences, some students may not turn on their cameras, making it challenging to determine their presence and participation in the online class. Consequently, it becomes difficult for teachers to observe the level of student engagement during online learning, especially during screen sharing to explain teaching materials. This situation presents a common obstacle in online learning. To address this obstacle, it is necessary to conduct research to develop methods of automatic students' engagement detection during the online learning process.

Students' engagement detection is an essential factor in improving the learning process. It is a qualitative indicator in the learning process [1]. It entails three structured learning dimensions: behavioral engagement, emotional engagement, and cognitive engagement [2]. While all the three dimensions of engagement are crucial for measuring students' level of involvement in the learning process, emotional engagement is the most widely studied. Detecting students' emotional engagement is particularly important in education because it has a significant impact on their learning rate and overall academic performance. Whitehill et al. [3] showed that both human and automatic engagement judgments are correlated with task performance. The study found that post-test student performance could be predicted based on engagement labels with similar accuracy to pre-test results.

The problem of automatically detecting students' engagement in online learning based on video data can be solved using a machine learning approach. Zang et al. [3] investigated engagement detection in online learning through a data-driven approach based on facial expressions and mouse usage behavior. Their study demonstrated that utilizing multiple features for detection could significantly improve the accuracy of engagement detection. In contrast to previous studies that solely relied on students' facial expressions, they also took into account students' mouse usage behavior in their approach. Bhardwaj et al. [4] proposed a deep learning model named Convolutional Neural Network (CNN) for students' engagement detection, while Selim, et al. [5] conducted students' engagement detection in online learning using Hybrid EfficientNetB7 together with TCN, LSTM, and Bi-LSTM. Khenkar et al. [6] also proposed an engagement detection method based on micro-body gestures using 3D Convolutional Neural Network (CNN).

Ashwin et al. [7] also conducted engagement detection using CCTV video recordings in a computer laboratory, in which case CCTV video recordings were successfully used to analyze students' engagement. Convolutional Neural Networks (CNN) were successfully implemented with a good level of accuracy in identifying students' engagement levels. This study's results revealed a positive correlation between students' scores (student learning) and students' predicted engagement levels. Meanwhile, Sharma et al. [8] detected students' engagement using video recordings of students' learning through emotional analysis and tracking of eye gaze and head movements based on two machine learning algorithms, namely the Haar Cascade algorithm (for face and eye detection) and the Convolutional Neural Network algorithm (CNN) (for emotion classification). Based on these studies, CNN is a powerful deep learning model that has been successfully used in various studies to detect students' engagement levels in online learning. By analyzing emotional features, tracking eye gaze directions, and estimating head movements, CNN could predict students' engagement levels, which is essential for improving the effectiveness of online learning.

One of the widely used datasets for video-based students' engagement detection is the DAiSEE dataset (Dataset for Affective States in E-Environments). The DAiSEE dataset was first introduced in the study of Gupta et al. in 2016 [9]. The benchmark accuracy value of the DAiSEE dataset for the affective level of engagement was 51.07%. Based on the benchmark evaluation result, there are still many opportunities for improving the classification performance of the DAiSEE dataset. The data distribution for each label of the affective level of engagement is unequal, with 1% for very low engagement, 5% for low engagement, 50% for high engagement, and 45% for very high engagement. This data imbalance can result in errors in the classification model, leading to overfitting or underfitting. One solution to address this issue is to balance the data using undersampling or oversampling techniques [10], [11]. Ali et al. [12] presented a data-level approach and an algorithm-level approach for handling class imbalance problems. Bach et al. [13] examined some undersampling and oversampling methods for highly imbalanced data. The conclusion of their research was that the Synthetic Minority Oversampling Technique (SMOTE) boosted by the Edited Nearest Neighbours (ENN) method allowed for an improvement in classification precision. Fernandez et al. [14] also revealed through their research that the SMOTE algorithm improved performance in supervised learning problems.

Therefore, imbalances in the DAiSEE dataset must be addressed. The current research's objective was to perform data balancing and feature selection to improve the benchmarking performance of the video-based students' engagement detection model on the DAiSEE dataset.

This article is organized as follows. Section II explains related works from previous studies. Section III describes the proposed model and methodology for students' engagement detection. Section IV presents the results of the methodology implementation. Section V provides a discussion of the results, and Section VI presents the conclusions of this study.

II. Related Works

Many studies related to the detection of students' engagement have been carried out. Bhardwaj et al. [4] used two datasets. The first one is the FER-2013 dataset, which is an image dataset used to train the CNN model, and the second one is the MES dataset, which is a tabular dataset used to do weight and subsequent calculations of the MES (Mean Engagement Score). The engagement level of students is classified into two classes: "engaged" and "not engaged." The proposed model achieved an accuracy level of 93.6%, a precision level of 98.48%, and a recall level of 87%. The proposed automated approach will certainly help educational institutions achieve an improved and innovative online learning method.

Selim et al. [5] also used the DAiSEE dataset to detect students' engagement and compared the performance of the proposed method with the VRESEE dataset. They proposed a Hybrid EfficientNetB7 model combined with TCN, LSTM, and Bi-LSTM. EfficientNet was pre-trained on the ImageNet dataset, which includes eight models ranging from EfficientNet B0 to EfficientNetB7. The study also compared the proposed and previous models on the DAiSEE dataset. The results of the three proposed models were as follows: EfficientNetB7+TCN, EfficientNetB7+Bi_LSTM, and EfficientNetB7+LSTM were at the levels of 64.67%, 67.39%, and 67.48%, respectively, outperforming the state-of-the-art ResNet+TCN model that was at 63.59%. When evaluating the proposed models on the VRESEE dataset, the highest accuracy achieved was 94.47% (from the use of EfficientNetB7+Bi_LSTM).

Paidja et al. [15] used the DAiSEE dataset for engagement emotion classification. They proposed a Convolutional Neural Network (CNN) model and performed feature extraction using five facial landmarks and the Euclidean distance between points and center points from the facial image. They also compared CNN with other machine learning algorithms, such as Support Vector Machine (SVM) and Deep Neural Network (DNN). The accuracy results obtained indicated that CNN successfully recognized engagement emotions better than the other methods. However, the limitation of their research was that it did not use the entire DAiSEE dataset as only 77 out of 9068 videos were used.

Abedi et al. [16] described the improvement of the state-of-the-art technology for detecting students' engagement using a ResNet and TCN Hybrid Network. This research also used the DAiSEE dataset and evaluated the performance of the ResNet+TCN method, comparing it to several previous studies on the DAiSEE dataset. The experimental results showed that the proposed ResNet+TCN model could improve the classification accuracy performance by 63.9%. It is very challenging to detect the minority engagement level with a very small sample in a supervised classification problem.

Zhang et al. [17] proposed an Inflated 3D Convolutional Network (I3D) for automatic students' engagement. The research also used the DAiSEE dataset for students' engagement detection, coupled with the use of OpenFace and AlphaPose for feature extraction. The proposed method achieved an accuracy of 52.35%.

Bajaj [18] et al. proposed SOTA hybrid ResNet+TCN for the detection of students' affective states. They also used the DAiSEE dataset with ResNet for feature extraction and TCN (Temporal Convolutional Network) for classification. The accuracy level reached by the study was 53.6%. The biggest challenge posed by this dataset is high class data imbalance.

Liao et al. [19] used the DAiSEE dataset and presented the Deep Facial Spatiotemporal Network (DFSTN) model for engagement prediction. To extract facial spatial features, they utilized pre-trained SE-ResNet-50 (SENet). The experiment obtained an accuracy of 58.84%.

Hasnine et al. [20] examined the extraction and visualization of students' emotions for engagement detection in online learning. The proposed model for emotion extraction and engagement detection consists of several steps. First, the OpenCV Face Recognition is implemented to detect emotions and eyes. This step results in emotion weight and eye gaze weight. These weights are used to calculate the Concentration Index (CI), which is then used to determine a student's engagement level based on specific rules. If the CI is greater than 65%, the student is detected to be highly engaged. If the CI is between 25% and 65%, the student is considered to be engaged. Otherwise, if the CI is less than 25%, the student is detected to be disengaged.

Brenner et al. [21] presented a social robot system that could detect a person's engagement by utilizing proxemics, body posture, and attention features. The proposed model achieved precision, recall, and F1 score results of 0.81, 0.82, and 0.81, respectively. The intended use of the proposed system is to design robots whose behaviors indicate awareness of a person's engagement.

Previous research commonly used video recording data to detect students' engagement. The DAiSEE dataset is one of the most popular datasets used in previous studies [5], [9], [15]–[19], [22]. Other datasets that have been used for this purpose include the EmotiW 2018 dataset [23], the EmotiW 2020 dataset [24], the Engagement Recognition (ER) database [25], the UPNA head pose dataset [26], and other video recordings [3], [7], [8], [20], [27].

The limitations of previously discussed DAiSEE dataset studies are related to model performance. The performance of detection models such as one with the DAiSEE dataset improved in accuracy from an average benchmark accuracy of 57.9% in 2016 [9] for baseline benchmarking, to 63.9% in 2020 [26], to 67.48% in 2022 [5]. The levels of accuracy are still relatively low, however, so there remain many challenges to overcoming this problem. In addition, the selection of the features to be extracted to increase the model accuracy still needs improvement.

III. METHODOLOGY

The research methodology used is illustrated in Fig. 1. It is important to note that the facial images appearing in this discussion were taken from the DAiSEE open dataset developed by [9].

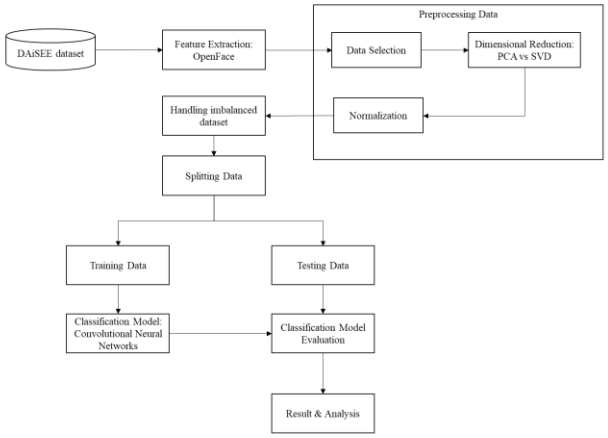

Fig. 1. Research methodology.

The explanation for each stage in the research methodology is as follows:

A. Dataset

The dataset used in this study was composed of secondary data from an open dataset called DAiSEE (Dataset for Affective States in E-Environments). The dataset was downloaded from https://people.iith.ac.in/vineethnb/resources/daisee/index.html. DAiSEE is a multi-label video classification dataset comprising 9,068 recorded video clips from 112 students, aimed at identifying students' affective states, including boredom, confusion, engagement, and frustration. Each affective state is labeled into four levels: very low, low, high, and very high. The videos were annotated by psychology experts and a crowd. This study focused solely on engagement levels, which were denoted by numbers: 0 for very low, 1 for low, 2 for high, and 3 for very high.

B. Feature Extraction

The subsequent step, feature extraction, was carried out using the OpenFace library. This open-source library is widely used for face recognition purposes, with the capabilities of facial landmark detection, head pose estimation, facial expressions (facial action units) recognition, and eye gaze estimation.

C. Data Pre-Processing

The next stage, data pre-processing, aimed to prepare data for the modeling stage. This stage involved three steps: data selection, feature dimensional reduction, and data normalization. The outcomes of this stage were feature matrices that could be utilized in the subsequent stage.

D. Imbalanced Dataset Handling

The oversampling or undersampling techniques could be employed to address data imbalances, which could lead to prediction errors in the model. Undersampling aimed to balance the data by reducing the number of instances in the majority class to match the number in the minority class. On the other hand, oversampling balanced the data by increasing the instances in the minority class to match those in the majority class.

E. Data Splitting

The feature matrix with balanced data was used in the training and testing processes of the classification model. The training data were used to form the classification model, while the testing data were used to evaluate the performance of the model formulated.

F. Classification Model

Fig. 2 is an illustration of the classification model formulation process. The video recording data collected with feature extraction were used as input data in the CNN classification model. The output of the CNN classification model was the prediction class or engagement level of the input video data.

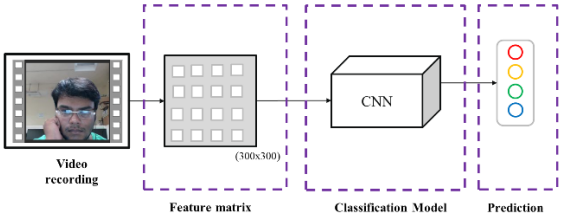

Fig. 2. The classification model formulation process.

G. Classification Model Evaluation

After obtaining the classification model through the training process, the model was tested using the testing data. The test results were then evaluated using metrics such as accuracy, precision, recall, and F1-score. These values were used to determine the performance of the proposed method. To further evaluate the classification model, the confusion matrix was referred to.

The confusion matrix in Fig. 3 is a matrix visualization of the prediction number and the actual data number on the classification model used. True positive (TP) is the number of correctly predicted data in the positive class. False positive (FP) is the number of incorrectly predicted data in the positive class. True negative (TN) is the number of correctly predicted data in the negative class. False negative (FN) is the number of incorrectly predicted data in the negative class [28].

		Prediction	
	Class	Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Fig. 3. Confusion matrix in binary classes.

The accuracy evaluation value compares the number of correctly predicted data with the entire data being tested. It can be calculated using Equation 1 based on the confusion matrix in Fig. 3.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

The precision evaluation value compares the number of correctly predicted data in the positive class with the overall positive predicted results. It can be calculated using Equation 2.

$$Precision\ (Pc) = (TP) / (TP + FP) \quad (2)$$

re

The recall or sensitivity evaluation value compares the number of correctly predicted data in the positive class with all the actual data in the positive class. It can be calculated using Equation 3.

$$Recall\ (Rc) = (TP) / (TP + FN) \quad (3)$$

Meanwhile, the F1-score evaluation value compares the average weighted precision and recall. It is better in measuring a classification model's performance than the precision or recall value. It can be calculated using Equation 4.

$$F1\text{-}Score = 2 * [(Pc * Rc) / (Pc + Rc)] \quad (4)$$

IV. Results

A. Dataset

The current study used an open dataset named DAiSEE, which is a video dataset that recognizes students' affective levels, including engagement. Each video has a clip ID and engagement level label: very low, low, high, or very high. Fig. 4 shows some examples of the downloaded DAiSEE dataset.

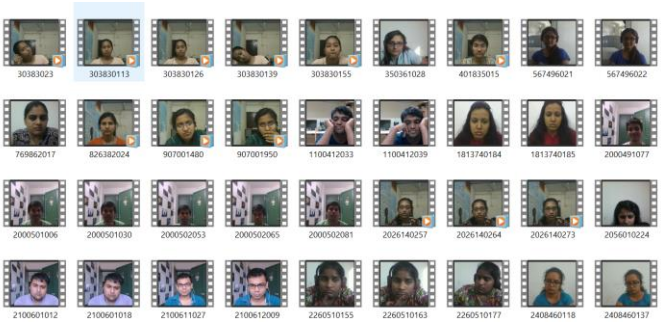

Fig. 4. Some examples of the downloaded DAiSEE dataset.

The data distribution for each engagement level can be seen in Table I. According to the table, the data for each level of engagement were highly imbalanced. Low and very low engagement levels were minority classes with data presentation of 0.7% and 5.1% of the total available data, respectively. If data of this sort are processed, it will cause errors in the classification model due to overfitting. This can be addressed by balancing the data with undersampling or oversampling techniques.

TABLE I. Distribution of Data on the DAiSEE Dataset for Each Level of Engagement

Engagement Level	Number of Videos	Percentages
0 (very low)	61	0.7%
1 (low)	455	5.1%
2 (high)	4422	49.5%
3 (very high)	3987	44.7%
Total	8925	100%

B. Feature Extraction

The OpenFace library extracted facial features from each video frame. It can be downloaded on the following GitHub page: https://github.com/TadasBaltrusaitis/OpenFace. Fig. 5 is an example of video output generated from the OpenFace library.

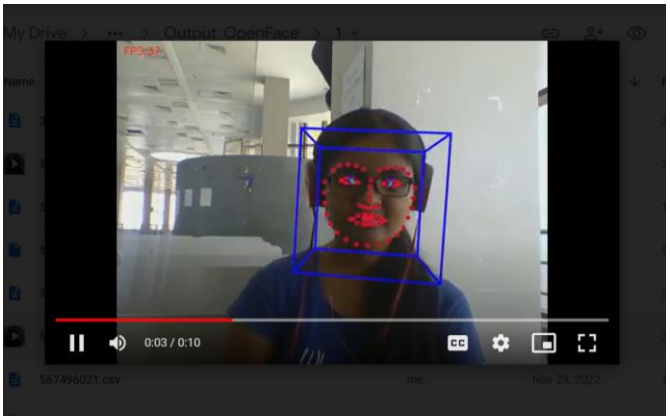

Fig. 5. An example of video output generated from the openface library.

In addition to the video output above, each video generated a CSV file. The file would display the columns frame, face ID, timestamp, confidence, success, and 709 facial feature values covering facial landmark detection, head pose estimation, eye gaze estimation, and estimation of facial expressions in the forms of facial action unit (AUs) features. The CSV file was also modified to store data of file name and the level of engagement for each frame. The DAiSEE dataset comprises 10-second videos with a frame rate of 30 fps, producing 300 frames for each video.

C. Data Pre-Processing

As shown in Fig. 1, there were three pre-processing stages: data selection, dimensional reduction, and data normalization.

1) Data selection: The first data selection stage involved selection of videos to facilitate the computational process. The video selection process was carried out in the following sub-stages:

- Find id_people from the video name (taken from the first five digits)
- Search for unique id_people
- Count and sort in the ascending the number of videos for each unique id_people
- Add up the cumulative value to the threshold = 61 (the total value of the minority class)
- Choose a video name based on the selected unique id_people

The results of feature extraction using the OpenFace library had a confidence column. This column refers to the confidence level of the model as to whether the detected face_id was a face or not. The confidence value ranged from 0 to 1. The closer it was to 1, the more confident the model was that the detected object was a face. On the other hand, the closer it was to 0, the less confident the model was that the detected object was a face.

	A	B	C	D	
1	frame	face_id	timestamp	confidence	su
344	65	0	2.133	0.03	
345	65	1	2.133	0.98	
346	66	0	2.167	0.03	
347	66	1	2.167	0.98	

Fig. 6. An example of a frame that detected two face objects.

As shown in Fig. 6, in frame 65, two objects, face_id 0 and face_id 1, were detected at the timestamp of 2.133 seconds. face_id 0 was detected at a confidence level of 0.03, and face_id 1 was at 0.98. If more than one object was found in a frame, data selection would be performed, where the object with the highest confidence value was to be selected. The data distribution for each engagement level before and after data selection in stages 1 and 2 can be seen in Table II.

TABLE II. DATA DISTRIBUTION BEFORE AND AFTER THE DATA SELECTION PROCESS

Engagement Level	Imbalanced Data		Stage 1 Data Selection		Stage 2 Data Selection	
	A	*B*	*A*	*B*	*A*	*B*
0 (very low)	61	0.7%	61	22.6%	59	23.4%
1 (low)	455	5.1%	63	23.3%	56	22.2%
2 (high)	4422	49.5%	70	25.9%	64	25.4%
3 (very high)	3987	44.7%	76	28.1%	73	29.0%
Total	8925	100%	270	100%	252	100%

A = Number of Videos, B = Percentages

2) Dimensional reduction: The length of the feature vector generated for each frame was very large, i.e., 1x709. Therefore, it became necessary to reduce the dimensions of the features to obtain unique features that could be used as differentiators for each level of engagement. The algorithms used at this stage were PCA (Principal Component Analysis) and SVD (Singular Value Decomposition). The explained variance value refers to the percentage value of the variance from the initial data. The number of components extracted covered a minimum of 80% of the explained variance in the data. In other words, at least 80% of the variance of the data was successfully captured. The greater the value of the explained variance, the better the original data were represented. Based on Table III, component = 300 was chosen because it had the highest explained variance value for both PCA and SVD. In addition, it was chosen so that each video produced would form a feature matrix with a square size of 300 x 300. Thus, using PCA and SVD, the number of features was reduced from 709 to 300.

TABLE III. THE EXPLAINED VALUES OF PCA AND SVD

Number of Components	PCA	SVD
2	81.91727	72.52122
3	96.99637	90.64553
10	99.80054	99.77922
50	99.99682	99.99676
100	99.99963	99.99962
200	99.99996	99.99996
300	99.99998	99.99998

3) Normalization: The feature matrices produced in the dimensional reduction stage had different ranges of values. It became necessary to normalize the data to prevent them from turning into noise in the model training process. The data normalization method used in this independent study was the min-max normalization method. This normalization method produced new feature values that had the same range from 0 to 1.

D. Imbalanced Dataset Handling

Based on Table II, the data for each level of engagement needed to be more balanced. It was necessary to balance the data to avoid overfitting prediction results. SMOTE (Synthetic Minority Over-sampling Technique), which synthesizes new data by re-sampling the minority class data to balance the data to the majority class, was used as an oversampling technique. A comparison was made between the number of data before and after applying SMOTE (see in Table IV).

TABLE IV. THE NUMBER OF DATA BEFORE AND AFTER SMOTE APPLICATION

Engagement Level	Before SMOTE	After SMOTE
0 (very low)	59	73
1 (low)	56	73
2 (high)	64	73
3 (very high)	73	73
Total	252	292

E. Data Splitting

Before entering the training stage of the classification model, the pre-processed data were divided into training and testing datasets, with 80% of the data being used for training and the remaining 20% for testing. The training data were used to form a supervised learning classification model, while the testing data were used to evaluate the classification model. The distribution of training and testing data for each engagement level can be found in Table V.

TABLE V. THE NUMBER OF TRAINING AND TESTING DATA

Engagement Level	Training Data	Testing Data	Total
0 (very low)	58	15	73
1 (low)	59	14	73
2 (high)	58	15	73
3 (very high)	58	15	73
Total Videos	233	59	292

F. Classification Model

The CNN classification model consisted of two stages, namely feature extraction and classification. The former consisted of four convolutional and pooling layer combinations as can be seen in Fig. 7. The feature maps on convolutional layer 1, layer 2, layer 3, and layer 4 were 32, 64, 128, and 256, respectively. The kernel size was 5 with the activation function using "ReLU". The latter used max-pooling with a pooling size of 2 x 2.

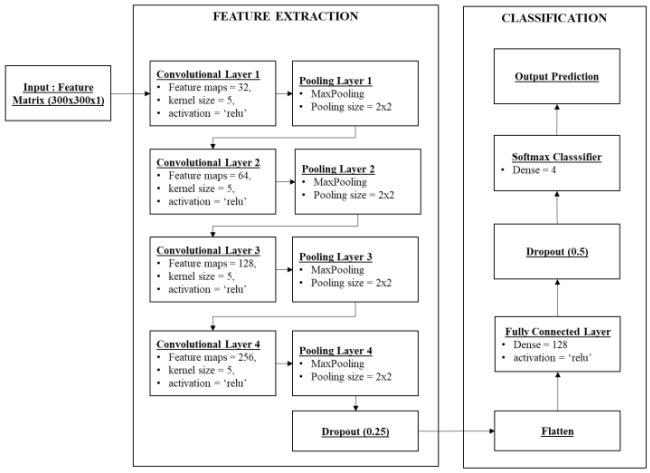

Fig. 7. Classification model of CNN.

The learning parameters used during the model-building process were batch size, epoch, learning rate, and optimizer. The trial-error approach was used to make the parameter selection. Table VI details the parameters of the CNN model used.

TABLE VI. THE AMOUNT OF TRAINING AND TESTING DATA

Parameter	Parameter Values
Number of epochs	800, 1600
Optimizer	Adam
Batch size	32, 16, 8, 4, 2
Learning rate	10^{-5}, 10^{-4}

In learning with artificial neural networks, the best model is often not found in the most recent epoch. Therefore, checkpoints and early stopping are used in the training process. A checkpoint is a CNN model that records each time the loss value decreases by a specified difference. In this way, if the loss value tends to increase or stagnate, the CNN model that manages to achieve the lowest loss value will be stored. Early stopping is a technique to stop the CNN learning process when the loss value has not shown a significant decrease in the number of certain epochs or when the model is said to have converged. This method is used because it can optimize the maximum number of epochs but saves more training time by stopping CNN training when it shows no improvement in learning. In early stopping, there is the patience parameter (p), which is used to determine the conditions for stopping training when it is found that the number of epochs remains the same as there is no decrease in the loss value. The patience value used in this independent study was half the number of epochs.

V. DISCUSSION

Based on the previous discussion, dimensional reduction was carried out with two approaches: PCA and SVD. Therefore, in evaluating this classification model, a comparison was made between the classification models from PCA-reduced data and SVD-reduced data.

Table VII shows the best evaluation value for each experiment. It can be seen that PCA-CNN had the highest accuracy of 72.88% in model 19, with an average accuracy value of 69.66% and parameter values as follows: optimizer = Adam, epoch = 1600, learning rate = 10-4, and batch size = 4. In comparison, model 8 had a higher maximum accuracy of 74.58% but with a standard deviation value greater than that of model 19 (3.54 > 3.34). The smaller standard deviation value was chosen because it means that the accuracy value in the experiment was closer to the average value.

TABLE VII. THE BEST EVALUATION VALUE FOR EACH EXPERIMENT

Parameter/Evaluation	**PCA**				**SVD**			
Model	**2**	**8**	**14**	**19**	**25**	**28**	**35**	**38**
optimizer	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam
Epoch	800	800	1600	1600	800	800	1600	1600
Learning rate	10-5	10-4	10-5	10-4	10-5	10-4	10-5	10-4
Batch Size	16	4	4	4	2	8	2	8
Average Accuracy	60.64	68.14	60.51	69.66	61.02	70.34	64.41	71.02
Standard Deviation	2.76	3.54	4.01	3.34	6.61	3.58	7.19	3.17
Minimum Accuracy	55.93	59.32	55.93	69.66	52.54	64.41	55.93	67.8
Maximum Accuracy	64.41	74.58	67.8	72.88	74.58	76.27	77.97	77.97

For experiments using SVD-CNN, the highest accuracy value was found in model 38, with a maximum accuracy value of 77.97% and an average accuracy value of 71.02%. The best parameter values obtained in this model were as follows: optimizer = Adam, epoch = 1600, learning rate = 10-4, and batch size = 8. Model 38 was found to have the smallest standard deviation value. This model was quite stable in providing accuracy evaluation values from the ten iterations performed for each model.

Regarding the learning rate parameter, the SVD-CNN and PCA-CNN experiments both had a learning rate of 10-4, producing the best accuracy model. Compared to the learning rate of 10-5, the learning rate of 10-4 provided faster computation time because the lower the learning rate, the higher the accuracy of the network, which means that the training process takes longer. For epoch size, the SVD-CNN and PCA-CNN experiments had the same number of epochs, 1600, which produced the best accuracy model. As can be seen in Table VII, when the epoch number of 800 was applied, the optimal accuracy value had yet to be reached. However, in terms of computational time, the larger the epoch number, the greater the time required. If we look at the batch size parameter, the SVD-CNN and PCA-CNN experiments had different parameter values generated by their best models. SVD-CNN had the best batch size of 8, while PCA-CNN had the best batch size of 4. The smaller the batch size, the more batches will be generated, requiring greater computation time. Table VII shows that SVD-CNN was better than PCA-CNN in terms of accuracy, batch size, number of epochs, and learning rate and required shorter computation time.

TABLE VIII. COMPARISON OF PRECISION, RECALL, AND F1-SCORE FOR EACH EXPERIMENT

			Engagement Level				
Experiment	*Model*	*Evaluation*	*0 (very low)*	*1 (low)*	*2 (high)*	*3 (very high)*	*Average*
PCA-CNN	Model 19	Precision	0.62	0.73	0.92	0.73	0.75
		Recall	0.87	0.57	0.73	0.73	0.73
		F1-Score	0.72	0.64	0.81	0.73	0.73
SVD-CNN	Model 38	Precision	0.85	0.83	0.81	0.67	0.79
		Recall	0.73	0.71	0.87	0.80	0.78
		F1-Score	0.79	0.77	0.84	0.73	0.78

Table VIII presents the comparison of precision, recall, and F1-score between PCA-CNN and SVD-CNN experiments. The two best models, models 19 and 38, had precision values higher than the recall values. This indicates false negatives or prediction errors in the actual engagement level data. Meanwhile, the F1-score indicates the average comparison value of weighted precision and recall. In model 19, engagement level 2 (high engagement) had the highest F1-score, i.e., 0.81. This high F1-score indicates that the classification model had fairly good precision and recall values.

Based on the accuracy, precision, recall, and average F1-score values, SVD-CNN performed better than PCA-CNN. Not only were they used for dimensional reduction, SVD and CNN were also used to select important features from the overall features. If analyzed from the variance value generated at the data pre-processing stage, PCA-CNN and SVD-CNN had the same variance value at component value = 300. The higher the variance value, the better the data representation to obtain unique information from the data. Meanwhile, if analyzed from the correlation value between features and engagement level, SVD-CNN had a higher correlation value than PCA-CNN. The features obtained from the SVD results had the best correlation value with the engagement level data. Therefore, in this study, SVD-CNN was superior to PCA-CNN.

The comparison of the values achieved by previous models and those by the model proposed in the DAiSEE dataset can be seen in Table IX. The PCA-CNN and SVD-CNN models with data balancing produced the highest accuracy performance at 72.88 and 77.97, respectively, with fewer features than the previous models.

TABLE IX. COMPARISON OF THE ACCURACY VALUES OF PREVIOUS MODELS WITH THE PROPOSED MODELS IN THE DAISEE DATASET

Model	**Feature (per frame)**	**Feature Dimensions**	**Accuracy**
I3D (Inflated 3D Convolutional Network) [17]	OpenFace (1x600) and AlphaPose (1x36) with feature selection	Not mention	52.35%
SOTA hybrid ResNet+TCN [16]	ResNet	1x512	53.6%
LRCN (Long-term Recurrent Convolutional Networks) [9] **– baseline benchmarking on DAiSEE**	**Not mention**	**Not mention**	**57.9%**
DFSTN (Deep Facial Spatiotemporal Network) [19]	SE-ResNet-50 (SENet)	1x2048	58.84%
ResNet + TCN [16]	ResNet	1x512	63.90%
3D DenseAttNet (DenseNet self-attention neural network) [22]	DenseAttNet	224x224x3	63.59%
EfficientNet B7 + LSTM [5]	EfficientNet B7	1x2560	67.48%
PCA-CNN with balanced data (proposed model)	OpenFace (1x709) with dimensional reduction PCA	1x300	72.88%
SVD-CNN with balanced data (proposed model)	OpenFace (1x709) with dimensional reduction SVD	1x300	77.97%

VI. CONCLUSION

In this study, we have successfully improved the benchmark performance of the DAiSEE dataset. The DAiSEE dataset experienced improvements from an average benchmark accuracy of 57.9% in 2016 [9] for baseline benchmarking, to 63.9% in 2020 [26], to 67.48% in 2022 [5]. We applied data balancing using oversampling and undersampling in the Convolutional Neural Network (CNN) classification model. The DAiSEE dataset also went through the pre-processing stages of data selection, dimensional reduction, and normalization. The features used in this study were taken from the OpenFace library, including 709 facial feature values from facial landmark detection, head pose estimation, eye gaze estimation, and facial expressions (facial action units (AUs)) estimations. Dimensional reduction was performed on the OpenFace features obtained using PCA and SVD techniques. A component number of 300 was applied in the PCA and SVD dimensional reduction, which means that the number of unique features was reduced from 709 to 300.

The PCA-CNN model had the highest accuracy rate of 72.88%, and the SVD-CNN model did 77.97%. The best CNN model parameter values were as follows: learning rate = 10-4, optimizer = Adam, epoch = 1600, and batch size = 4 (PCA-CNN) and 8 (SVD-CNN). The two PCA-CNN and SVD-CNN

best models had precision values higher than the recall values (0.75 > 0.73 for PCA-CNN and 0.79 > 0.78 for SVD-CNN). This indicates that there were false negative events such as prediction errors in the actual engagement level data. Meanwhile, the highest F1-score values were 0.73 (PCA-CNN) and 0.78 (SVD-CNN), which shows that the classification models had fairly good precision and recall values.

From all the experiments that have been carried out, it can be concluded that SVD-CNN had better performance than PCA-CNN in evaluating average, maximum, precision, recall, and F1-score values accuracy. If analyzed from the variance value generated at the data pre-processing stage, PCA-CNN and SVD-CNN had the same variance value at component value = 300. The higher the variance value, the better the data representation to obtain unique information from the data. Meanwhile, if analyzed from the correlation value between features and engagement level, SVD-CNN had a higher correlation value than PCA-CNN. Moreover, it can also be interpreted that the features obtained from the SVD results had the best correlation value with the level of engagement contained in the data. Therefore, in the current study, SVD-CNN was superior to PCA-CNN.

Although this study has provided better evaluation results than previous studies on the DAiSEE dataset, there remains a room for improvement for further research. It is necessary to explore alternative approaches to determining the optimal component values to produce features that have a more significant impact on the classification model. Additionally, conducting a more in-depth analysis of features beyond facial expressions can increase the accuracy of students' engagement detection.

ACKNOWLEDGMENT

This research was partially supported by an internal publishing grant from the Faculty of Computer Science, University of Indonesia.

REFERENCES

[1] F. D'Errico, M. Paciello, and L. Cerniglia, "When emotions enhance students' engagement in e-learning processes," Article in Journal of E-Learning and Knowledge Society, vol. 12, no. 4, pp. 9–23, 2016, doi: 10.20368/1971-8829/1144.

[2] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "School Engagement: Potential of the Concept, State of the Evidence," 2004.

[3] Z. Zhang, Z. Li, H. Liu, T. Cao, and S. Liu, "Data-driven Online Learning Engagement Detection via Facial Expression and Mouse Behavior Recognition Technology," Journal of Educational Computing Research, vol. 58, no. 1, pp. 63–86, Mar. 2020, doi: 10.1177/0735633119825575.

[4] P. Bhardwaj, P. K. Gupta, H. Panwar, M. K. Siddiqui, R. Morales-Menendez, and A. Bhaik, "Application of Deep Learning on Student Engagement in e-learning environments," Computers and Electrical Engineering, vol. 93, Jul. 2021, doi: 10.1016/j.compeleceng.2021.107277.

[5] T. Selim, I. Elkabani, and 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, pp. 99573–99583, Sep. 2022, doi: 10.1109/access.2022.3206779.

[6] S. Khenkar and S. K. Jarraya, "Engagement detection based on analyzing micro body gestures using 3D CNN," Computers, Materials and Continua, vol. 70, no. 2, pp. 2655–2677, 2022, doi: 10.32604/cmc.2022.019152.

[7] T. S. Ashwin and R. M. R. Guddeti, "Unobtrusive students' engagement analysis in computer science laboratory using deep learning techniques," in Proceedings - IEEE 18th International Conference on Advanced Learning Technologies, ICALT 2018, Aug. 2018, pp. 436–440. doi: 10.1109/ICALT.2018.00110.

[8] P. Sharma, S. Joshi, S. Gautam, S. Maharjan, V. Filipe, and M. C. Reis, "Student Engagement Detection Using Emotion Analysis, Eye Tracking and Head Movement with Machine Learning."

[9] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards User Engagement Recognition in the Wild," Sep. 2016, [Online]. Available: http://arxiv.org/abs/1609.01885.

[10] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," ACM Computing Surveys, vol. 52, no. 4. Association for Computing Machinery, Aug. 01, 2019. doi: 10.1145/3343440.

[11] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," GESTS International Transactions on Computer Science and Engineering, vol. 30, 2006.

[12] A. Ali, S. Mariyam Shamsuddin, A. Ralescu, and A. L. Ralescu, "Classification with class imbalance problem: A review," Classification Int. J. Advance Soft Compu. Appl, vol. 5, no. 3, 2013, [Online]. Available: https://www.researchgate.net/publication/288228469.

[13] M. Bach, A. Werner, J. Żywiec, and W. Pluskiewicz, "The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis," Inf Sci (N Y), vol. 384, pp. 174–190, Apr. 2017, doi: 10.1016/j.ins.2016.09.038.

[14] A. Fernández, S. García, F. Herrera, and N. V Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," 2018.

[15] A. N. R. Paidja and F. A. Bachtiar, "Engagement Emotion Classification through Facial Landmark Using Convolutional Neural Network," in Proceedings - 2022 2nd International Conference on Information Technology and Education, ICIT and E 2022, 2022, pp. 234–239. doi: 10.1109/ICITE54466.2022.9759546.

[16] A. Abedi and S. S. Khan, "Improving state-of-the-art in Detecting Student Engagement with Resnet and TCN Hybrid Network," in 18th Conference on Robots and Vision (CRV), Apr. 2021, pp. 151–157. [Online]. Available: http://arxiv.org/abs/2104.10122.

[17] H. Zhang, X. Xiao, T. Huang, S. Liu, Y. Xia, and J. Li, "An Novel End-to-end Network for Automatic Student Engagement Recognition," in IEEE7539th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC), 2019, pp. 342–346.

[18] K. K. Bajaj, I. Ghergulescu, and A.-N. Moldovan, "Classification of Student Affective States in Online Learning using Neural Networks," Nov. 2022, pp. 1–6. doi: 10.1109/smap56125.2022.9942163.

[19] J. Liao, Y. Liang, and J. Pan, "Deep facial spatiotemporal network for engagement prediction in online learning," Applied Intelligence, vol. 51, no. 10, pp. 6609–6621, Oct. 2021, doi: 10.1007/s10489-020-02139-8.

[20] M. N. Hasnine, H. T. T. Bui, T. T. T. Tran, H. T. Nguyen, G. Akçapõnar, and H. Ueda, "Students' emotion extraction and visualization for engagement detection in online learning," in 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, Procedia Computer Science, 2021, vol. 192, pp. 3423–3431. doi: 10.1016/j.procs.2021.09.115.

[21] M. Brenner, H. Brock, A. Stiegler, and R. Gomez, "Developing an engagement-aware system for the detection of unfocused interaction," in 2021 30th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2021, Aug. 2021, pp. 798–805. doi: 10.1109/RO-MAN50785.2021.9515353.

[22] N. K. Mehta, S. S. Prasad, S. Saurav, R. Saini, and S. Singh, "Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement," Applied Intelligence, vol. 52, no. 12, pp. 13803–13823, Sep. 2022, doi: 10.1007/s10489-022-03200-4.

[23] Y.-Y. Li and Y.-P. Hung, "Feature Fusion of Face and Body for Engagement Intensity Detection," in 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 3312–3316.

[24] O. Copur, M. Nakıp, S. Scardapane, and J. Slowack, "Engagement Detection with Multi-Task Training in E-Learning Environments," Apr. 2022, [Online]. Available: http://arxiv.org/abs/2204.04020.

[25] O. M. Nezami, M. Dras, L. Hamey, D. Richards, S. Wan, and C. Paris, "Automatic Recognition of Student Engagement using Deep Learning and Facial Expression," CoRR, vol. abs/1808.02324, Aug. 2018, [Online]. Available: http://arxiv.org/abs/1808.02324.

[26] S. S. Mane and A. R. Surve, "Engagement Detection using Video-based Estimation of Head Movement," in 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT-2018), 2018, pp. 1745–1749.

[27] K. Altuwairqi, S. K. Jarraya, A. Allinjawi, and M. Hammami, "A new emotion–based affective model to detect student's engagement," Journal of King Saud University - Computer and Information Sciences, vol. 33, no. 1, pp. 99–109, Jan. 2021, doi: 10.1016/j.jksuci.2018.12.008.

[28] M. Grandini, E. Bagli, and G. Visani, "Metrics for Multi-Class Classification: an Overview," Aug. 2020, [Online]. Available: http://arxiv.org/abs/2008.05756.