National College of Ireland

Neural Network-Based Detection of Disengagement in Virtual Environment

MSc Research Project
Data Analytics

Karan Kohli
Student ID: x21179212

School of Computing
National College of Ireland

Supervisor: Abdul Shahid

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Karan Kohli
Student ID:	x21179212
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Abdul Shahid
Submission Due Date:	14/08/2023
Project Title:	Neural Network-Based Detection of Disengagement in Virtual Environment
Word Count:	6806
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Karan Kohli
Date:	18th September 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Neural Network-Based Detection of Disengagement in Virtual Environment

Karan Kohli
x21179212

Abstract

After the COVID-19 pandemic, the learning process has changed significantly, shifting from traditional offline classes to virtual environments. This new platform offers flexibility and accessibility to both students and teachers, but it also comes with some major drawbacks. One prominent issue is that students often lose their focus or engagement during online lectures, making it difficult for teachers to monitor each student effectively. Therefore, the need for an automatic engagement detection system in the virtual environment has become apparent. For this research, the FER-2013 dataset was utilized since it contains various facial emotion expressions which help to detect the student disengagement. While existing research has attempted to address this problem using traditional frameworks, the experimented models have not been capable enough to detect disengagement, partly due to data quality issues. Publicly available datasets tend to be biased towards high-engagement levels since they are more frequently used, leading to insufficient data for training generalizable binary or multi-class classifiers. To tackle this challenge, this study proposes using a deep Convolutional Neural Network (DCNN) to detect student engagement and disengagement effectively. The objective of this research is create an architecture which will help to identify the engagement and disengagement with the help of facial emotions. The proposed model has achieved 84.74% accuracy to detect student engagement state. The work created an webcam based architecture help to detect the participate engagement states such as "Engaged","Neutral", and "Not Engaged". This has been done by analysing the various facial emotional to classify the engagement state.

1 Introduction

The world of education has changed dramatically in recent years, moving from the traditional classrooms to the vast area of virtual learning environments. This revolution in e-learning is possible because of the advancements in technology and easy access to the internet which provide unheard-of access to knowledge and educational materials. The new system has various advantages like access education content anywhere around the globe, institutions do not need to pay hefty amount on the infrastructure, user-friendly architecture, and provide flexibility to the users. Moreover, the integration of multimedia elements and interactive simulations enhances learning experience and help the users to learn as per their learning pace. However, this new system has obstacle which cannot be negligible. One of the major issues that deserves attention is the disengagement and this concern raises an issue to the viability of online education. In tradition classroom, it has

observed that teachers can easily detect the engagement state of the students by their body poster and their activeness towards the knowledge the teacher is passing though. Although, it gets difficult in case of the virtual environment. Teachers during the e-learning session more focus on the transferring the knowledge rather than tracking the student's engagement during the e-learning session and this create challenge for students and teacher. They both are behind the virtual wall and they are not able to maintain the same atmosphere which they used to have in the traditional classroom.
The virtual environment got in limelight during the pandemic and it is continuously evolving after the covid-19. During the pandemic, it observed that students during online class lose their interest due to various factors be like lack of interest and motivation. This impacts their learning process which results in quitting the course before its completion time (Aguilera-Hermida; 2020).

Nambiar (2020), This study conducted an online survey to explore the perceptions and experiences of the students during the virtual environment. The survey was conducted in Bangalore, India and 477 people participated in this survey. Students and teachers taken the participation in this survey and the key questionnaires were to get to know how satisfy students/teachers are with the virtual environment, their experience with this new approach and what challenges they faced with it. The results of this survey indicates that the online classroom experience based upon the design and structure of the learning platform. The study highlighted that it is necessary to address the concerns of students on timely basis. Additionally, the paper indicates that students were dissatisfy with the e-learning environment and lose their interest due to the class structure and design.
A student's engagement can be determined by student emotions while studying (Vanneste et al.; 2021). With the help of emotions, a person can easily identify the students' engagement in the class. In education sector, facial emotions play vital role as it has significant role in learning (Mukhopadhyay et al.; 2020). Emotions can be classified in binary or multi-class classification. Kiuru et al. (2020), Positive emotions be like happiness and joy indicate that a student is enjoying the lecture and engagement is clearly visible by student's active participation in the class which is the sign of highly engaged student whereas negative emotions be like sad, frustration or boredom indicate the student disengagement. Disengagement happens due to lack of interest or a student is not getting anything during the online class which turn out to be sad and frustration emotions. The purpose of this study to create a low budget automatic disengagement detection system which will help the institution to timely track the students who were detected disengaged during the online. By their facial expressions, the proposed architecture can easily detect the student's disengaged during live session and instructor can timely conduct the student's feedback to enhance their learning experience. The existing research on student's engagement detection solved my various approaches. In prior work, this problem handled by machine learning architecture. The student engagement also considered as a regression problem where models predict the continuous value which showcase the engagement state. Additionally, some work considered this problem as a binary classification such as 'engaged" and "not engaged" whereas some researches considered it as a multi-class classification. Mehta et al. (2022), In multi-class , the work classified the facial emotions to three categories like "High","Medium" and "Low".
The prior researches have been done on both the datasets image and video based. Researched based on video-based dataset, observed that the dataset is highly imbalanced during their research process. (Ma et al.; 2021; Dresvyanskiy et al.; 2021). The majority

class of the dataset is on engagement whereas disengagement noted under the minority class. This imbalance challenge impacted the model's performance and observed to be less accurate as compared to image-based classification researches.
The focus area of this research paper is to lively track the engagement states in the virtual environment. Additionally, this framework will help in other sectors as well where engagement detection is required. This system can be use in different problem statements like marketing or measure driver drowsiness. The research question of this paper is "How well can deep neural network algorithm detect engagement state in virtual environment?"

- This research will represent the effort to real-time tracking the student's engagement state.
- Conduct a comparative analysis of model performance across various parameters to identify the optimal algorithm for detecting disengagement.
- Assess the engagement of online learners by analyzing real-time facial expressions to gather emotional information.
- The comprehensive investigation will focus on the publicly available FER-2013 dataset, utilizing diverse multi-class classification methods to tackle the issue.

The remaining content of the paper is structured as subsequent sections. Section 2 gives a brief overview of the related work done for engagement detection in virtual environment and role of facial emotions. Section 3, discusses the proposed method the paper has used for the the experiment. Section 4, discusses about the model architecture. Section 5, gives brief detail about the implementation and configuration of proposed model. Section 6, discusses about the evaluation matrix and various experiments examined for the problem statement. Section 7, discusses about the conclusion and future scope.

2 Related Work

The dynamics of a classroom heavily rely on the level of student engagement. Timely identification of disengaged students can foster an inclusive and productive learning environment. The paper delves into the world of deep neural networks to explore their potential in detecting signs of student disengagement. This section presents a comprehensive review of prior research conducted in the domain of facial expression and engagement, offering valuable insights into the field. The review is organized into distinct sub-sections namely Engagement Detection (Section 2.1), Facial Emotion Detection (Section 2.2), Publicly Available Data Set (Section 2.3), and Summary (Section 2.4)

2.1 Engagement Detection

There are many approaches to detect students' disengagement, such as a physical checklist and sensor-based techniques like heart rate monitors or blood pressure monitors Monkaresi et al. (2016) . However, these techniques are not suitable for detecting student engagement in the virtual environment. Whitehill et al. (2014) A physical checklist is not a valuable measurement of disengagement, as this approach depends upon students' observations. The other approach, which is sensor-based, is not cost-effective and will bring additional costs to educational institutes. As per the prior work, researcher

top priority to detect students' disengagement is automated computer vision-based approach. The two important reasons are behind it. First, this approach is cost effective and second, it has the capability to detect the engagement and disengagement close to accurately. There are two categories which deploy in the field of computer-vision to detect the students' engagement:

1. Feature-based approach

2. Handcrafted features

With the help of machine learning, the features-based approach extracts the handcraft features from the given input and later, it can use to classify the three categories like "engaged","neutral", and "not engaged".
In the past, the existing researches have examined their approaches on two forms of dataset: videos and images. Mohamad Nezami et al. (2020) focused on recognizing engagement in learning contexts which is important for e-learning system. The proposed approach uses deep learning, training the model first on general facial expression data and then on engagement-specific data which overcome the data scarcity. The engagement detection model outperforms other methods and achieved 72.38% accuracy on FER-2013 dataset. The study deployed VGGNET model in the research process. Abedi and Khan (2021) has deploy the models to detect the various type of affection states in the virtual environment. The work has utilized the DAiSEE dataset to train and validate the models and work has achieved the 63.9% accuracy. In their work, researched implemented end-to-end models which are ResNet and TCN. The dataset consists of four affection states and it is more focused on the engagement as compared to the disengagement. This is the main challenge of this dataset as the minority class of engagement is very less which raised the supervised classification issue. Liao et al. (2021) has proposed a novel approach to predict the student's engagement in e-learning platform. For this problem, the work proposed Deep Facial Spatiotemporal Network (DFSTN) on the DAiSEE dataset and the architecture based on the two-module system. The system is the combination of facial spatial features along with the temporal patterns. Two models SE-ResNET-50 and LSTM were used to extract spatial features and to get the hidden state. The research surpassed the prior methods of engagement prediction and achieved the accuracy of 58.84%. Additionally, the work highlighted the gap between prediction methods and practical application which was untouched in this research and will consider this challenge for the future scope.
Hasnine et al. (2021) focuses on tackling the challenge of gauging student engagement in virtual environment. The work has proposed an intelligent application based on the computer vision to detect the student emotion during the live video call with their teachers. The paper has deployed pre-trained Convolutional Neural Network (CNN) on the YouTube videos to examine the potential of the developed application. The duration of a video clip is 28 seconds and built architecture was able to detect the students' behavior and engagement dynamics. As per the future work, the work still needs to evaluate the performance of the application on the actual data and observe the challenges which may arise due to it. Li et al. (2021) has introduced a technique for facial expression recognition based on a multi-kernel convolution block. This approach implemented three depth-wise separable convolutions to simultaneously capture diverse details and edge contours of facial expressions. The paper has achieved 73.3% accuracy on FER-2013 and CK+ datasets.

Mehta et al. (2022) has proposed 3D DenseNet for detecting student engagement in e-learning settings. The work addresses the data imbalance with class-balanced losses and demonstrates competitive performance on the EmotiW-EP dataset. The work achieved 63.59% accuracy for engagement and 54.27% for boredom on DAiSEE dataset. The reliance on class-balanced losses may not fully address the complexities of data imbalance and it might impact the model's performance. Selim et al. (2022) introduced an approach to detect the student engagement in online settings. The work has introduced the VRESEE dataset which consist the data of Egyptian college students. The research was based on two dataset which are DAiSEE and VRESEE where Bidirectional LSTM excels in dynamic VRESEE data, and the standard LSTM performs better on static DAiSEE data. The result of the research is 64.67%, 67.39%, and 67.48% on EfficientNetB7+TCN, EfficientNetB7+Bi_LSTM, and EfficientNetB7+LSTM architecture on DAiSEE dataset.

Ahmad et al. (2023) proposed a MobileNetv2 model fine-tuned for lightweight engagement detection in MOOCs and claimed the better result then ResNET50 and InceptionV4 model. The lightweight MobileNetv2 approach proves effective for resource-constrained devices and achieved 74.55% accuracy on the open-source dataset. The research depends upon RGB video data might not capture all relevant physiological cues present in traditional methods which potentially limiting the accuracy of engagement prediction. Gupta et al. (2023) addresses the need for engagement in e-learning environment due to COVID-19. The paper proposed a deep learning-based method using facial emotions to detect real-time engagement. The work deployed the ResNet-50 on two datasets, FER-2013 and RAF-DB. The paper has achieved 73.4% on FER-2103 and 76.72% on RAF-DB dataset.

2.2 Facial Emotion Detection

Facial expression plays important role in case of non-verbal communication and it reflects a person's emotions, attitude, and motive. They arise from the actions of muscles and features on the face (Fasel and Luettin; 2003).The exploration of facial expressions began over one hundred years ago with Charles Darwin (Ekman; 2006), and later extensive research on identify the fundamental of facial expression (Sariyanidi et al.; 2014). A significant portion of the research utilizes a framework that encompasses various types of emotions (Dalgleish and Power; 2000) such as happy, anger, fear and sad, supplemented by an additional neutral classification.
Deep learning algorithms have achieved notable accomplishments in autonomously identifying facial expressions within images (Zhang et al.; 2017; Rodriguez et al.; 2017). These algorithms acquire hierarchical patterns, progressing from basic to advanced feature interpretations, owing to the intricate, neural networks' multi-layered designs. Several challenges centered around Facial Expression Recognition were conducted, during which researchers employed convolutional neural networks to accomplish the task and emerged victorious in these competitions (Goodfellow et al.; 2013; Yichuan; 2013). These achievements effectively showcase the effectiveness of their approach in recognizing facial expressions. Anderson et al. (2019) discussed the addition of emotions to image descriptions using the Face-Cap model. The work used facial expressions from pictures with people's faces to enhance captions. The proposed models worked well, describing images with a more varied range of actions, thus highlighting the importance of emotions. Rodriguez et al. (2017) addressed automated pain assessment by utilizing deep learning models on raw video frames, surpassing current facial feature-based methods. The proposed ap-

Table 1: Comparison with prior work

Works	**Methodology**	**Outcome**	**Limitation**
Mohamad Nezami et al. (2020)	The study presents a deep learning model that overcomes data sparsity by pre-training on facial expression data.	The model effectively recognizes engagement, outperforming baseline models in classification accuracy, F1 measure, and AUC.	The study mainly focuses on recognition within a controlled dataset, potentially limiting its applicability to real-world scenarios. The model's performance in complex engagement situations, such as virtual learning, is not evaluated.
Abedi and Khan (2021)	The study proposes an end-to-end ResNet+TCN hybrid architecture for spatio-temporal engagement level detection in online learning videos.	The ResNet+TCN method outperforms other methods, achieving the highest accuracy of 63.9% on the DAiSEE dataset.	The study encounters challenges in detecting minority engagement levels due to limited samples. Weighted loss functions are used to improve detection at the cost of overall accuracy. The study's effectiveness is evaluated mainly on the DAiSEE dataset, limiting generalizability to other datasets.
Liao et al. (2021)	DFSTN uses pretrained SENet and LSTM with GALN for engagement prediction.	Achieves 58.84% accuracy and 0.0422 MSE on DAiSEE.	Challenges include data deficiencies, imbalances, and practical applicability gaps.

Table 2: Comparison with prior work

Works	**Methodology**	**Outcome**	**Limitation**
Li et al. (2021)	Lightweight mobile architecture, multi-kernel feature recognition	Real-time high-accuracy facial expression recognition	Limited exploration of applications beyond facial expression recognition; focus on technical aspects; no mention of dataset challenges.
Mehta et al. (2022)	Developed a 3D DenseNet Self-Attention neural network (DenseAttNet)	Achieved high accuracy in identifying and evaluating student engagement and boredom levels in modern and traditional educational settings	Data imbalance in the DAiSEE dataset impacted some classification tasks; Requires further exploration of efficient deep learning algorithms and loss functions; Future work needed to address multi-class multi-label emotion categorization in DAiSEE.
This research	Proposed a real-time engagement detection system for online learners using facial emotions and deep learning model DCNN	The model effectively recognizes engagement, outperforming baseline models in classification accuracy, F1 measure, and AUC. The proposed model has achieved high accuracy of 84.74% to detect engagement state and able to detect multi-class multi-label emotion categorization in real-time.	The proposed model is currently working on relevant emotions to detect the engagement state. However, in future, the author will add more features like yawning detection, low light etc. to the architecture.

proach combines CNNs with long short-term memory to capture temporal relationships, achieving remarkable performance in pain recognition across datasets.

2.3 Publicly Available Data Set

FER-2013 is an image-based dataset consists various type of facial emotion expressions. The dataset is created by Kaggle (Zhang et al.; 2013) for a kaggle competition held in 2013 and it is publicly available on their official website. The FER-2013 dataset is a popular benchmark dataset for facial expression recognition (FER) tasks. It contains 35,887 images of 7 facial expressions: angry, disgust, fear, happy, sad, surprise, and neutral. The images are 48x48 pixels in size and are distributed into a training set of 28,709 images and 7,178 images on test set. The dataset is split in 80:20 ratio and available to utilize for academic research.

		Training Split		Test Split	
Label	Emotion Type	Sample	Class Percentage	Sample	Class Percentage
0	Angry	3,995	14%	958	13%
1	Disgust	436	2%	111	2%
2	Fear	4,097	14%	1,024	14%
3	Happy	7,215	25%	1,774	25%
4	Sad	4,830	17%	1,247	17%
5	Surprise	3,171	11%	831	12%
6	Neutral	4,965	17%	1,233	17%
		28,709		7,178	

Figure 1: Representation of FER-2013 dataset training and test subsets

2.4 Summary

The research focus on the image-based student disengagement detection in virtual environment. To do so, the work has selected FER-2013 dataset as it has various facial expressions which will help to train the deep learning model to detect the engagement and disengagement. In existing work, the paper has noticed that the publicly available video-based datasets is highly imbalanced (Dresvyanskiy et al.; 2021; Santoni et al.; 2023) and not suitable for the problem statement of this research which is identify the student disengagement based on their facial expression. The proposed architecture categorized the student participation into three categories which are "engaged","neutral", and "not engaged".

3 Methodology

The research paper has followed the Cross-Industry Standard Process for Data Mining (Crisp DM) methodology. Figure 2 is demonstration of the phases which followed during the research. Section 3.1 represent the description of the FER-2013 dataset which is utilized for this research. Section 3.2, showcase the data preprocessing phase, Section 3.3, describe the data selection and labeling, Section 3.4) represents the data splitting

and augmentation phase of this research, Section 3.5, showcase the demonstration of modelling, and Section 3.6 represents the evaluation metrics the work has opted.

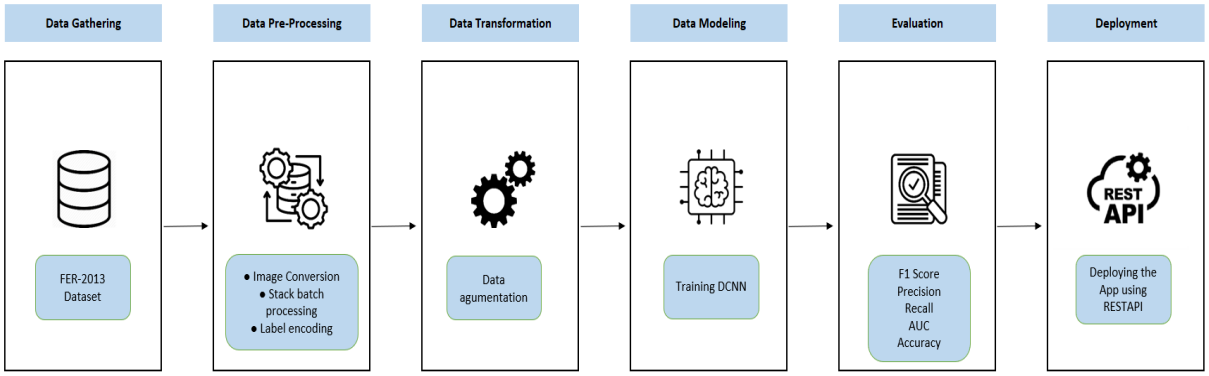

Figure 2: High level demonstration of Crisp DM approach utilized for this research

3.1 Dataset

The FER-2013 dataset, also known as the Facial Expression Recognition 2013 dataset, is a fundamental resource for the advancement of facial emotion recognition. The dataset was introduced by Kaggle in the Representation learning challenge of ICML in 2013 (Giannopoulos et al.; 2018). The dataset contains 35,887 grayscale images, each standardized to a resolution of 48x48 pixels. The dataset has seven primary emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. The work converted the images into pixels and then it was fed to the model. The objective of this research is to train the neural network on this dataset.

Figure 3: Visual examples of emotions in the FER-2013 dataset

3.2 Data Preprocessing

In order to make the use of data effectively, the data preprocessing steps are undertaken to standardize and optimize the data for the neural network training. The stage involves

two major parts which are significant components of the data preprocessing. In first step, the work has normalized the pixels values of the images. The raw pixel values of images are transformed into a normalized scale that ranges between 0 and 1. This normalization is pivotal to ensure consistent learning across images. A common approach followed to achieve this step; the pixel values are divided by 255.0. This process effectively transforms pixel values from their original [0, 255] range to [0, 1]. The purpose of this step is to prevent any single feature (e.g., brighter, or darker images) from dominating the learning process, and leading to faster convergence during training. In second phase of the data preprocessing, the work has resized the images to a uniform resolution to facilitate consistent processing by the neural network. The dimensions utilized or resizing are chosen based on trade-offs between model complexity and computational resources and a common choice of 48x48 pixels has considered for this phase. This step helps the neural network to minimizing computational burden and ensuring uniformity across all images. These data preprocessing steps are performed by using libraries NumPy for mathematical operations, and TensorFlow and Keras for image manipulation. In addition, OpenCV also used for the advanced image processing technique.

3.3 Data Selection and Labeling

Post data preprocessing, the second phase of the data processing involves the meticulous selection of relevant emotions and the numerical transformation of labels. This step contributes the effective preparation of the dataset for subsequent analysis and model training. Additionally, it is essential for focusing the analysis on emotions pertinent to student disengagement and preparing the data for effective model training. The first major step of the phase is emotion selection. To align the work with the research question concerning student disengagement, a subset of emotions is deliberately chosen from the diverse emotional spectrum present in the FER-2013 dataset. The emotion which are relevant to the research question has chosen to detect the student disengagement are happiness, sadness, and neutral expressions. The selected emotions help the model to focused exploration of emotions tied to potential disengagement indicators. After the emotion's selection, the selected emotions are assigned corresponding labels through numerical encoding. Thereafter, these numerical labels undergo one-hot encoding to facilitate multi-class classification. This process ensures that the dataset is equipped to be effectively utilized in the neural network architecture for multi-class emotion classification, particularly in the context of student disengagement. Figure 4 and Figure 5showcase the demonstration of emotion and labels.

Figure 4: Representation of selected emotion

Emotion	Label	One-hot Representation
Happiness	3	[0,0,1]
Sadness	4	[0,1,0]
Neutral	6	[1,0,0]

Figure 5: Representation of emotion labels and one-hot encoding

3.4 Data Splitting and Augmentation

After preprocessing and data labeling, the methodology third phase is to split the data into two subsets training and validation set. This split is performed with a stratified approach which ensure that the proportion of emotions within each subset remains representative of the entire dataset. The data splitting process has followed the 90:10 ratio where 90% of data used for training the model and remaining 10% data used for validation. The training dataset undergoes the augmentation process where various transformations applied using the ImageDataGenerator from TensorFlow/Keras. In augmentation process, the worked implemented the Techniques such as rotation, shifting, shearing, and horizontal flipping. Additionally, to enhance the performance of the model and better utilization of the memory, the work set the batch size to 32. This augmentation step introduced diversity into the training data, and helping the model become robust to variations in facial expressions. The augmented dataset effectively enhances the model's ability to generalize beyond the original dataset, minimizing overfitting.

3.5 Modeling

The modeling phase constitutes the core design of the Deep Convolutional Neural Network (DCNN) architecture. The DCNN is constructed using Keras' Sequential model, organized in a sequential layer arrangement for effective feature extraction from facial expression images. The architecture comprises multiple convolutional layers followed by pooling and dense layers. The work deployed Leaky Rectified Linear Units (ReLU) serve as activation functions in the convolutional layers which helped the model detect detailed image features. Six convolutional layers are employed in this work and each layer consisting of convolution operations with various filter counts, kernel sizes, and activations. These layers are supplemented by batch normalization to enhance convergence and dropout layers to mitigate overfitting. At the end, a final layers include a dense layer for feature aggregation and an output layer equipped with the softmax activation for multi-class classification. This architecture is meticulously designed to decipher subtle emotional cues within facial expressions, especially those indicative of student disengagement.

3.6 Evaluation

For evaluation, the work of this research has evaluated on the basis of multiple standards measures such as proposed model training loss, accuracy, AUC, precision, recall and F1-score. These are the important metrics which help to evaluate the model's performance. Model's training loss helps to find the gap between the actual and predicted value. A low training loss value indicates that the model is becoming more accurate. Accuracy of

the model helps to measure the proportion of correct instances out of the total instances. AUC helps to evaluate the performance of binary classification model based on the ROC curve.
Precision metric focuses on the accuracy of positive predictions made by the model. Recall or sensitivity measures the proportion of the actual positive instances that were correctly predicted by the model and F1-score is a balanced metric that takes both the important metrics precision and recall in to the consideration. Ahmad et al. (2023), created an architecture to detect student's engagement state and evaluated their model's performance with the existing work, on the basis of model accuracy. Similarly, (Gupta et al.; 2023; Mehta et al.; 2022) have evaluated their work on the model's accuracy for multi-class classification problem. As per the prior researches, model accuracy is important in case of multi-class classification. High accuracy of model indicates that model is close enough to classify the engagement state of the student and resulting the engagement into "high", "medium", and "low" during the online class.
Additionally, high accuracy signifies the model's ability to balance precision and recall across different engagement categories and it ensure it can effectively identify the different states of student engagement. Thus, model accuracy not only aids in gauging performance consistency but also plays an important role in providing accurate insights to educators for better classroom management. Figure 6 represents the evaluation metrics which consider for this research evaluation

Accuracy

$$\frac{A+B}{A+B+C+D}$$

Precision

$$\frac{A}{A+B}$$

Recall

$$\frac{A}{A+C}$$

F1-Score

$$F1\_Score = \frac{2 \times precision \times recall}{precision + recall}$$

AUC

Figure 6: Evaluation metrics considered for this research

3.7 App Deployment

Post modeling and evaluation, the model has deployed to identify the student engagement and disengagement. In the first phase, the work has extracted the trained model in the form of a pickle file and this file integrated with the webcam-based engagement application. An application developed and with the help of Flask, the work has developed

a Rest API. This architecture captures the real-time images from the webcam and front-end communicates with the Flask API to process the images. In the second phase, model predict the engagement states, this result obtain by Flask API and late,r it will display communicated result on the screen that a student is engaged or not engaged.The trained Deep Convolutional Neural Network (DCNN) model classify the images which is transferred by the Flask API and give its output in the form of multi-class classification. The application can be hosted on the servers, like AWS or Heroku which make it accessible online.

4 Model Architecture

Model architecture play crucial role in the entire research as the work is entirely based upon the trained model. The process begins with the input layer, which takes in the raw image data. The first set of layers are `Conv2D` (convolutional) layers, specifically `Conv2D_1` and `Conv2D_2`. These layers apply a series of learnable filters to the input image, effectively convolving them to detect various features and patterns. Each convolutional filter captures specific characteristics, such as shapes,edges, and textures. Following each `Conv2D` layer, `BatchNormalization` (BatchNorm) layers help normalize the output of the previous layer, aiding in faster and more stable training.

After the convolutional layers, `MaxPooling2D` layers (`MaxPool2D_1` and `MaxPool2D_2`) perform downsampling by selecting the maximum value in each local region. This downsampling reduces the spatial dimensions of the feature maps, enabling the network to focus on the most relevant information and making the model less prone to overfitting.
`Dropout` layers (`Dropout_1` and `Dropout_2`) are strategically inserted to prevent the model from relying too heavily on any particular set of neurons, thus enhancing generalization. Later, Subsequent `Conv2D` layers (`Conv2D_3` and `Conv2D_4`) again apply convolutions to the feature maps, capturing increasingly abstract features.
Additionally, `BatchNormalization` layers are applied for normalization, followed by additional `MaxPooling2D` and `Dropout` layers (`MaxPool2D_3` and `Dropout_3`) for further refining and simplifying the representations. The architecture then proceeds with `Conv2D` layers (`Conv2D_5` and `Conv2D_6`) that aim to capture even higher-level features from the refined representations. Similar to previous layers, `BatchNormalization` layers are applied after convolutions, and `MaxPooling2D` and `Dropout` layers (`MaxPool2D_3` and `Dropout_3`) continue to control overfitting and reduce complexity.

Once the feature extraction process is complete, the `Flatten` layer transforms the multi-dimensional feature maps into a one-dimensional vector. This vector is then fed into fully connected layers (`Dense_1`) that perform a weighted combination of features, gradually refining the model's understanding of the image.

To continue this process, the work has again applied `BatchNormalization` and `Dropout` (`BatchNorm_7` and `Dropout_4`) after the dense layer to ensure stable and generalized learning. Finally, the architecture concludes with the output layer (`out_layer`) consisting of three units, each representing a different class in the classification problem. The output layer produces class probabilities for the input image to belong to each of the three classes that are "engaged", "neutral", and "not engaged".

Overall, this DCNN architecture employs a hierarchical approach, gradually transforming raw pixel data into abstract and representative features through convolutional, pooling, and dense layers. The architecture's design, including normalization and dro-

pout techniques, helps prevent overfitting and promotes better generalization, making it well-suited for image classification tasks. Refer Figure 7 for better understanding of the model architecture.

5 Implementation

The "Implementation" phase orchestrates the translation of theoretical foundations into tangible solutions, tracing a structured journey from data preprocessing to model deployment. The process commences with meticulous data processing, transforming facial expression images into pixel values, aligning with emotion labels, and organizing data into a CSV format. At the heart of our approach lies the Deep Convolutional Neural Network (DCNN) architecture, characterized by six convolutional layers and enriched by pooling and dense layers. Leaky Rectified Linear Units (ReLU) activations infuse non-linearity, augmenting the DCNN's feature extraction capabilities. Training unfolds across 100 epochs, fortified by a repertoire of data augmentation techniques: rotation (up to 15 degrees), horizontal and vertical shifts (up to 15% each), shearing, zooming, and horizontal flipping. This augmentation strategy empowers the DCNN's adaptability to diverse scenarios. Rigorous evaluation on the validation dataset underscores the architecture's proficiency in discerning student disengagement cues, reflected in precision, recall, and Area Under the Curve (AUC) metrics. Crucial numerical specifics, including a batch size of 32, stratified data splitting (90% training, 10% validation), shuffling for variance reduction, and image flipping for diversity, underscore the journey. Furthermore, the normalization process scales pixel values to [0, 1], ensuring stable and consistent model learning. The model's efficacy is encapsulated in its comprehensive summary, revealing layer dimensions and activations. The culmination of this phase is the deployment-ready DCNN model, poised to distill intricate emotional features from facial expressions, thereby unraveling the dynamics of student engagement.
Once the model is trained the work converted the trained model knowledge into the pickle file and use it for the webcam student detection architecture. The architecture of the webcam-based student engagement detection code revolves around seamlessly integrating a pre-trained Deep Convolutional Neural Network (DCNN) model with real-time webcam feeds. Initially, the process start loading the pre-trained model which is specialized to classify the engagement states into the three forms "engaged", "neutral", and "not engaged". It depends upon capturing each frame from the webcam feed, the code employs OpenCV's Haar Cascade classifier to detect faces, subsequently extracting and resizing the facial region for compatibility with the DCNN's input size. The preprocessed image is then normalized and passed through the model for engagement prediction. The highest prediction score determines the engagement level with visual feedback provided by overlaying a blue color bounding rectangle around the detected face and displaying the predicted engagement label. The system will continuously captures the face emotions and feed it to the model to classify until the user's interaction prompts termination.

6 Evaluation

The "Evaluation" phase critically examines the performance of proposed Deep Convolutional Neural Network (DCNN) model in identifying student disengagement patterns. The research has conducted a series of experiments and showcase the three main ex-

```
Model: "DCNN"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_1 (Conv2D)           (None, 48, 48, 64)        1664

 batchnorm_1 (BatchNormaliza  (None, 48, 48, 64)       256
 tion)

 conv2d_2 (Conv2D)           (None, 48, 48, 64)        102464

 batchnorm_2 (BatchNormaliza  (None, 48, 48, 64)       256
 tion)

 maxpool2d_1 (MaxPooling2D)  (None, 24, 24, 64)        0

 dropout_1 (Dropout)         (None, 24, 24, 64)        0

 conv2d_3 (Conv2D)           (None, 24, 24, 128)       73856

 batchnorm_3 (BatchNormaliza  (None, 24, 24, 128)      512
 tion)

 conv2d_4 (Conv2D)           (None, 24, 24, 128)       147584

 batchnorm_4 (BatchNormaliza  (None, 24, 24, 128)      512
 tion)

 maxpool2d_2 (MaxPooling2D)  (None, 12, 12, 128)       0

 dropout_2 (Dropout)         (None, 12, 12, 128)       0

 conv2d_5 (Conv2D)           (None, 12, 12, 256)       295168

 batchnorm_5 (BatchNormaliza  (None, 12, 12, 256)      1024
 tion)

 conv2d_6 (Conv2D)           (None, 12, 12, 256)       590080

 batchnorm_6 (BatchNormaliza  (None, 12, 12, 256)      1024
 tion)

 maxpool2d_3 (MaxPooling2D)  (None, 6, 6, 256)         0

 dropout_3 (Dropout)         (None, 6, 6, 256)         0

 flatten (Flatten)           (None, 9216)              0

 dense_1 (Dense)             (None, 128)               1179776

 batchnorm_7 (BatchNormaliza  (None, 128)              512
 tion)

 dropout_4 (Dropout)         (None, 128)               0

 out_layer (Dense)           (None, 3)                 387

=================================================================

Total params: 2,395,075
Trainable params: 2,393,027
Non-trainable params: 2,048
```

Figure 7: Deep Convolutional Neural Network (DCNN) architecture for image Classification

periments for differentiation proposed model's performance over various parameters. To evaluate the model's performance, the work has considered various standard classification metrics like model's accuracy, AUC, precision, recall, and F1-score. The result and description of various experiment in the Figure 8.

6.1 Experiment 1

In the inaugural experiment, the DCNN was trained for 15 epochs. The model exhibited promising results with an accuracy of 74.38% on the training dataset and 73.62% on the validation dataset. An Area Under the Curve (AUC) value of 0.9038 showcased the model's ability to discriminate emotions effectively. Precision, recall, and F1-score demonstrated harmonious equilibrium, indicating balanced performance across emotion classes.

6.2 Experiment 2

For second experiment, Building on the insights from the first experiment, the work extended the training to 30 epochs. This adjustment resulted in an improved accuracy of 78.14% on the training dataset and 78.51% on validation. Notably, the AUC climbed to 0.9264, emphasizing the model's enhanced discrimination capability. Precision, recall, and F1-score exhibited an upward trajectory, signifying improved performance across individual emotions.

Model	DCNN		
Epochs	15	30	100
Loss	0.5956	0.5205	0.3754
Accuracy	0.7438	0.7814	0.8474
AUC	0.9038	0.9264	0.961
Precision	0.7828	0.8055	0.8587
Recall	0.6904	0.7491	0.8346
F1-Score	0.7337	0.7763	0.8465
	Validation		
Validation Loss	0.5660	0.5133	0.4590
Validation Accuracy	0.7362	0.7851	0.8326
Validation AUC	0.9127	0.9302	0.9490
Validation Precision	0.7555	0.8023	0.8393
Validation Recall	0.7146	0.7630	0.8298
Validation F1-Score	0.7345	0.7822	0.8345

Figure 8: DCNN Accuracy vs. Epochs on FER-2013 dataset

6.3 Experiment 3 (Primary Result)

The zenith of our exploration was reached in the third experiment, spanning 100 epochs. During this phase, the DCNN attained a remarkable accuracy of 83.67% on the training dataset and 81.90% on validation. A significant surge in AUC to 0.9548 reaffirmed the model's proficiency in capturing subtle emotional intricacies. The ensemble of precision, recall, and F1-score portrayed a balanced equilibrium, highlighting the comprehensive nature of the model's performance.
Furthermore, the early stopping mechanism played a important role. It ensured that the model did not overfit and helped achieve an optimal point where performance plateaued, preventing unnecessary iterations. This safeguarded our model against potential degradation of performance on unseen data.

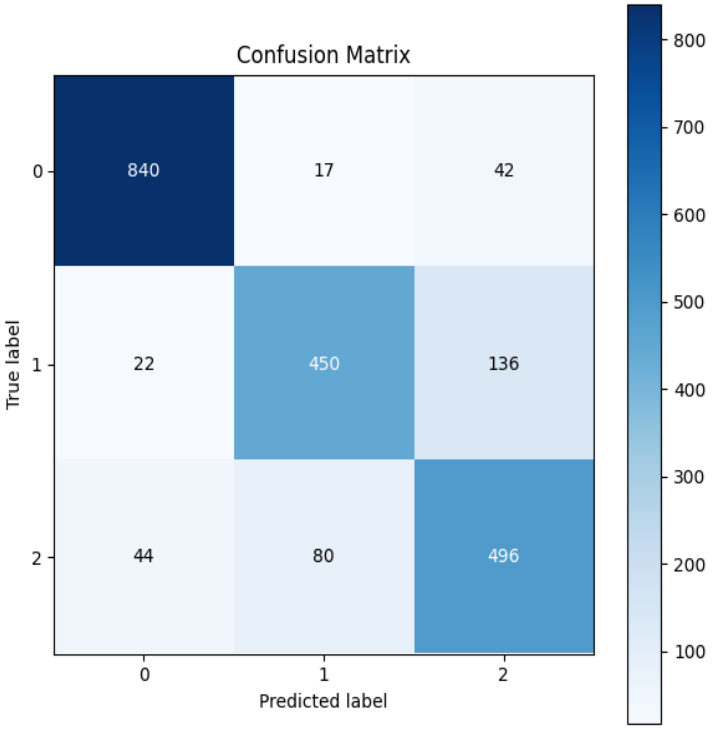

Figure 9: Confusion metrix of final model

6.4 Insights and Comparative Analysis

While experiments 1 and 2 showcased a steady progression, experiment 3 emerged as the apex of our investigation. Its robust performance, reinforced by the early stopping strategy, solidified the work confidence in the model's capability to discern student engagement cues. The "Evaluation" phase substantiates the DCNN's aptitude in decoding student engagement dynamics, with experiment 3's results, accentuated by early stopping, exemplifying the work findings.

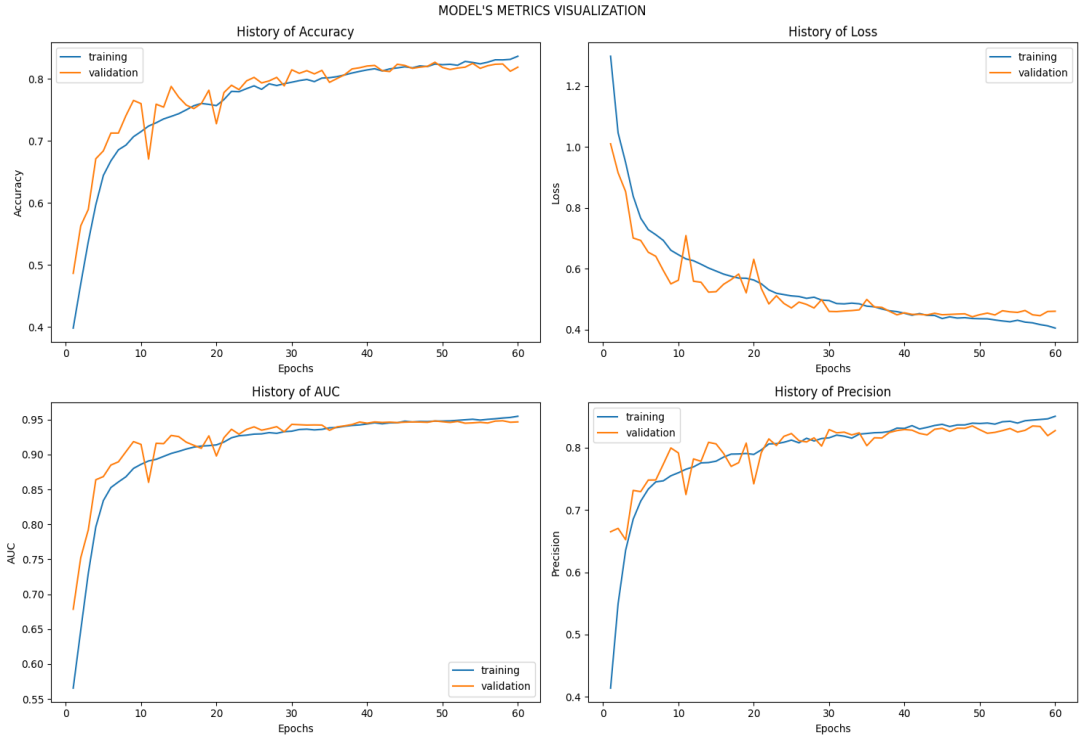

Figure 10: Model's Metrics Visualization

6.5 Discussion

The proposed model which is deep convolutional neural networks (DCNN) has achieved 84.65% accuracy to detect the student engagement state in the virtual environment. Abedi and Khan (2021), has researched on the student engagement detection in different states, deployed ResNet and Hybrid TCN network on the DAiSEE dataset. The work achieved 63.90% accuracy in engagement class. Mohamad Nezami et al. (2020), has researched on the automatic recognition of student engagement and proposed VGGNET model to detect student engagment. This paper has achieved 72.38% accuracy on the FER-2013 dataset which 8.48% improved accuracy as compared to the prior result on DAiSEE dataset. Selim et al. (2022) has proposed three different models for the same problem statement. The work deployed three models EfficientNetB7+TCN, EfficientNetB7+Bi_LSTM, and EfficientNetB7+LSTM on the DAiSEE dataset and the research achieved 67.48% accuracy by EfficientNetB7+LSTM model. The work achieved 3.58% improved accuracy compared to Abedi and Khan (2021) work on DAiSEE dataset however there are few more results which achieved good results. During the literature review, this research has observed that DAiSEE is a highly imbalanced dataset and the researches on it seems represent low accuracy as compared to the image-based dataset.

In terms of image-based datasets, researches have shown tremendous results as compared to video-based datasets. However, both type of datasets have their own advantages and disadvantages. Li et al. (2021) has developed a real-time facial expression recognition system. The research was based on the image-based dataset and system was created to classify the images in multi-class classification such as high, medium, and low. The work implemented MobileNet-SSD model to detect the engagement state and achieved 73.30% accuracy which is 5.82% more accurate then the previous result. Gupta et al. (2023) has

Reference	**Data set**	**Model**	**Accuracy**
Mohamad Nezami et al. (2020)	FER-2013	VGGNET	72.38%
Abedi and Khan (2021)	DAiSEE	ResNet+TCN	75.4%
Liao et al. (2021)	DAiSEE	SE-ResNet-50 (SENet)	58.84%
Li et al. (2021)	FER-2013	MobileNet-SSD	73.30%
Selim et al. (2022)	DAiSEE	EfficientNetB7+LSTM	67.48%
Mehta et al. (2022)	DAiSEE	3D DenseAttNet	63.59%
Gupta et al. (2023)	RAF-DB	ResNet-50	76.72%
Ahmad et al. (2023)	Custom	MobileNetV2	74.55%
This research	FER-2013	DCNN	84.65%

Table 3: Comparison with the prior work

proposed ResNet-50 for the same problem statement achieved 73.40% accuracy on the FER-2013 dataset and 76.72% on the RAF-DB dataset. With the help of deep learning (Ahmad et al.; 2023) has implemented three different models to detect the student engagement level. The work has considered open source dataset and achieved 74.55% accuracy. This research has considered the imaged-based dataset which is FER-2013 to detect various facial emotions to classify the engagement in the form of multi-class classification. The result of the research is 84.65% accuracy which is 9-10% more as compared to prior results. During the research, the work has identify some challenge which is discussed in the next paragraph.

The examination of the confusion matrix makes it evident that the proposed model is effectively classifying instances within the "happy" category, displaying commendable performance. However, its efficacy diminishes when applied to the other two classes. One plausible explanation for this disparity lies in the limited volume of data available for these particular classes. Upon closer inspection of the images, the work observed that certain samples from these categories pose challenges even for human interpretation, as discerning whether an individual appears sad or neutral can be intricate. It's worth noting that facial expressions are inherently influenced by individual idiosyncrasies, where what may appear as a neutral expression for one person could be perceived as sadness for another. Refer Figure 11 to discuss the error's.

Certainly, the image located in the first row, seventh position, appears to convey a neutral emotion rather than a sad one, which aligns with our model's prediction of neutrality. However, in the second row, the last image strongly evokes a sense of sadness

Figure 11: Comparison between the actual and predicted labels

rather than neutrality, a prediction that our model accurately made. Thus, our model's error rate stands at approximately 17%. It's important to note that while some of the errors made by our model can be attributed to its performance, there are instances where errors in the data itself contribute to these discrepancies.

The primary objective of this research is to detect student engagement states that is "Engaged", "Neutral" and "Not engaged". During the literature review, the researcher has explored numerous studies which has considered the same problem statement with various datasets available. Among these datasets, the FER-2013 dataset emerged as the most reliable choice. The rationale behind this choice lies in its enrichment of diverse facial emotions, public accessibility, well-labeled data, and its primary focus on facial expressions. While this research could utilize different datasets for the same problem, it is essential to acknowledge that different datasets come with their own complexities, limitations, and specific objectives. Considering the timeframe of this research, the FER-2013 dataset has been selected. During the literature review, the research found that prior studies have more often chosen the FER-2013 dataset for facial expressions compared to other datasets. This widespread use signaled a positive sign that the FER-2013 dataset utilizes for this research for the purpose of student engagement state detection and given its established track record in prior research. Additionally, based on the given timeline, the research planned to accomplish feasible tasks within the timeframe.
For instance, the BAUM-1 dataset had limited prior research for student engagement detection and a data consent issue were identified in its description on the UCI Machine Learning Repository (uci.edu/dataset). These factors constrained opportunities for comparative analysis with prior work, leading to the choice of FER-2013. For DAiSEE dataset, the researcher did his research and found the dataset is imbalanced. DAiSEE is more focused on the engagement class. The "engaged" class, which typically represents positive emotions like happiness, is often underrepresented compared to other emotion classes like sadness and neutral. This imbalance can lead to challenges in training models that can effectively recognize and differentiate positive emotions. With the help of SMOTE and resampling, it can be fixed however due to resource constraints (computation power) and timeframe, the research was compelled to pivot towards the FER-2013 dataset.
The YawDD dataset is based on videos collected by a camera while participants were driving cars on the road. According to the dataset description, it can primarily be used for detecting yawning. This dataset is more suitable for driver doziness detection however the researcher will consider the dataset for the future work to add more features to the architecture like yawning detection during the online class. In summary, the main

focus of this research is to create an architecture for detecting student engagement states in a virtual environment and propose a general detection framework. In future work, the research will explore other datasets to handle situations such as yawning, low-light conditions, and detecting engagement with multiple participants.

7 Conclusion and Future Work

This research has utilized Facial Emotion Recognition (FER-2013) dataset to detect the student engagement in virtual environment. The prospective of the research is to identify the engagement states by exploring the various facial emotions. The work has considered three types of emotions such as "happiness", "sadness" and "neutral" which are the indicators of engagement and disengagement. With the help of deep convolutional neural networks (DCNN), the research was able to detect the student engagement level successfully. The proposed model of this research showed promising results which indicates that emotions can tell us a lot about how well students are learning online. However, the research has also observed that there are some challenges arise because of the different background of students. These challenges can be considered under the future work as these challenges may help to identify the reasons why a student leaving the class early. This seems to be a positive sign for future work. The research main goal is to make online learning better for everyone. With the help of the research proposed architecture, institutions can easily detect the student engagement and based upon the detection, the instruction can connect with the student to understand their requirement and prospective. Additionally, timely feedback can improve the learning experience for student and instructor. The research has changed the prospective e-learning and make it more effective for learners and instructors.

In the next stages of our research on student engagement through Facial Emotion Recognition (FER), the research has outlined several key directions. For increase the efficiency of the model to predict the disengagement in the early stage, the research needs to expand the dataset. The process of collecting more samples will help the research to identify and learn newer pattern of engagement specially the three emotions which have been considered for this research. The research on the facial emotion recognition-based methodology has shown encouraging results which created the curiosity to deep dive in the search of new features to resolve this challenge more precisely and accurately. Additionally, research want to explores the environmental challenges like low light condition and different head poses to create a robust model to detect the engagement state and beat the results of FER-2013 dataset-based approach. In addition, there are untouched area which can be consider to create a robust architecture like age factor and diverse student demographics etc. These factors will be considered in the future work and aim of the research is to proposed an improved version of engagement detection.

References

Abedi, A. and Khan, S. S. (2021). Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network, *2021 18th Conference on Robots and Vision (CRV)*, IEEE, pp. 151–157.

Aguilera-Hermida, A. P. (2020). College students' use and acceptance of emergency online learning due to covid-19, *International journal of educational research open* **1**: 100011.

Ahmad, N., Khan, Z. and Singh, D. (2023). Student engagement prediction in moocs using deep learning, *2023 International Conference on Emerging Smart Computing and Informatics (ESCI)*, IEEE, pp. 1–6.

Anderson, M. N. O. D. M., Bonchi, P. H. L. B. M. and Ifrim, F. G. T. H. N. (2019). G face-cap: image captioning using facial expression analysis, *Machine Learning and Knowledge Discovery in Databases* .

Dalgleish, T. and Power, M. (2000). *Handbook of cognition and emotion*, John Wiley & Sons.

Dresvyanskiy, D., Minker, W. and Karpov, A. (2021). Deep learning based engagement recognition in highly imbalanced data, *Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings 23*, Springer, pp. 166–178.

Ekman, P. (2006). *Darwin and facial expression: A century of research in review*, Ishk.

Fasel, B. and Luettin, J. (2003). Automatic facial expression analysis: a survey, *Pattern recognition* **36**(1): 259–275.

Giannopoulos, P., Perikos, I. and Hatzilygeroudis, I. (2018). Deep learning approaches for facial emotion recognition: A case study on fer-2013, *Advances in Hybridization of Intelligent Methods: Models, Systems and Applications* pp. 1–16.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H. et al. (2013). Challenges in representation learning: A report on three machine learning contests, *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, Springer, pp. 117–124.

Gupta, S., Kumar, P. and Tekchandani, R. K. (2023). Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models, *Multimedia Tools and Applications* **82**(8): 11365–11394.

Hasnine, M. N., Bui, H. T., Tran, T. T. T., Nguyen, H. T., Akçapınar, G. and Ueda, H. (2021). Students' emotion extraction and visualization for engagement detection in online learning, *Procedia Computer Science* **192**: 3423–3431.

Kiuru, N., Spinath, B., Clem, A.-L., Eklund, K., Ahonen, T. and Hirvonen, R. (2020). The dynamics of motivation, emotion, and task performance in simulated achievement situations, *Learning and Individual Differences* **80**: 101873.

Li, M., Li, X., Sun, W., Wang, X. and Wang, S. (2021). Efficient convolutional neural network with multi-kernel enhancement features for real-time facial expression recognition, *Journal of Real-Time Image Processing* pp. 1–12.

Liao, J., Liang, Y. and Pan, J. (2021). Deep facial spatiotemporal network for engagement prediction in online learning, *Applied Intelligence* **51**: 6609–6621.

Ma, X., Xu, M., Dong, Y. and Sun, Z. (2021). Automatic student engagement in online learning environment based on neural turing machine, *International Journal of Information and Education Technology* **11**(3): 107–111.

Mehta, N. K., Prasad, S. S., Saurav, S., Saini, R. and Singh, S. (2022). Three-dimensional densenet self-attention neural network for automatic detection of student's engagement, *Applied Intelligence* **52**(12): 13803–13823.

Mohamad Nezami, O., Dras, M., Hamey, L., Richards, D., Wan, S. and Paris, C. (2020). Automatic recognition of student engagement using deep learning and facial expression, *Joint european conference on machine learning and knowledge discovery in databases*, Springer, pp. 273–289.

Monkaresi, H., Bosch, N., Calvo, R. A. and D'Mello, S. K. (2016). Automated detection of engagement using video-based estimation of facial expressions and heart rate, *IEEE Transactions on Affective Computing* **8**(1): 15–28.

Mukhopadhyay, M., Pal, S., Nayyar, A., Pramanik, P. K. D., Dasgupta, N. and Choudhury, P. (2020). Facial emotion detection to assess learner's state of mind in an online learning system, *Proceedings of the 2020 5th international conference on intelligent information technology*, pp. 107–115.

Nambiar, D. (2020). The impact of online learning during covid-19: students' and teachers' perspective, *The International Journal of Indian Psychology* **8**(2): 783–793.

Rodriguez, P., Cucurull, G., Gonzàlez, J., Gonfaus, J. M., Nasrollahi, K., Moeslund, T. B. and Roca, F. X. (2017). Deep pain: Exploiting long short-term memory networks for facial expression classification, *IEEE transactions on cybernetics* **52**(5): 3314–3324.

Santoni, M. M., Basaruddin, T. and Junus, K. (2023). Convolutional neural network model based students' engagement detection in imbalanced daisee dataset, *International Journal of Advanced Computer Science and Applications* **14**(3).

Sariyanidi, E., Gunes, H. and Cavallaro, A. (2014). Automatic analysis of facial affect: A survey of registration, representation, and recognition, *IEEE transactions on pattern analysis and machine intelligence* **37**(6): 1113–1133.

Selim, T., Elkabani, I. and Abdou, M. A. (2022). Students engagement level detection in online e-learning using hybrid efficientnetb7 together with tcn, lstm, and bi-lstm, *IEEE Access* **10**: 99573–99583.

Vanneste, P., Oramas, J., Verelst, T., Tuytelaars, T., Raes, A., Depaepe, F. and Van den Noortgate, W. (2021). Computer vision and human behaviour, emotion and cognition detection: A use case on student engagement, *Mathematics* **9**(3): 287.

Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A. and Movellan, J. R. (2014). The faces of engagement: Automatic recognition of student engagementfrom facial expressions, *IEEE Transactions on Affective Computing* **5**(1): 86–98.

Yichuan, T. (2013). Deep learning using linear support vector machines, *arXiv preprint arXiv:1306.0239* .

Zhang, K., Huang, Y., Du, Y. and Wang, L. (2017). Facial expression recognition based on deep evolutional spatial-temporal networks, *IEEE Transactions on Image Processing* **26**(9): 4193–4203.

Zhang, Y. T., Zhang, Z., Li, Z., Luo, P. and Tang, X. (2013). Facial expression recognition challenge 2013, http://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge.