Indian Journal of Information Sources and Services
ISSN: 2231-6094 (P) Vol.16, No.2, 2026, pp.379-394

DOI: https://doi.org/10.51983/ijiss-2026.16.2.38

Development of a Multimodal Student Engagement Level Prediction Dataset Using ISED and Classroom Video Data

Renuka Tantry[1], Dr. Sharada U Shenoy[2*], Dr. Shreenath Acharya[3], Dr.R. Rashmi Adyapady[4] and K.N. Prathibha[5]

[1]Research Scholar, NMAM Institute of Technology (NMAMIT), NITTE (Deemed to be University), Karkala, Karnataka, India; Assistant Professor, Department of Artificial Intelligence and Machine Learning, St Joseph Engineering College, Mangaluru, Karnataka, India
[2*]Professor & Head, Department of Artificial Intelligence and Machine Learning, NMAM Institute of Technology (NMAMIT), Nitte (Deemed to be University), Karkala, Karnataka, India
[3]Professor & Head, Department of Computer Science and Engineering (CSE-ICB), AJ Institute of Engineering & Technology (AJIET), Mangaluru, Karnataka, India
[4]Assistant Professor, Department of Artificial Intelligence and Machine Learning, NMAM Institute of Technology (NMAMIT), NITTE (Deemed to be University), Karkala, Karnataka, India
[5]Assistant Professor, Department of Mathematics, St Joseph Engineering College, Mangaluru, Karnataka, India
E-mail:[1]renuka.22phdecs208@student.nitte.edu.in; renukat@sjec.ac.in, [2]sharadauday@nitte.edu.in, [3]shree.katapady@gmail.com, [4]rashmi.adyapady@nitte.edu.in, [5]prathibhak@sjec.ac.in
ORCID: [1]https://orcid.org/0000-0002-8531-1275, [2]https://orcid.org/0000-0002-5947-3318, [3]https://orcid.org/0000-0003-3834-0003, [4]https://orcid.org/0000-0001-6370-3385, [5]https://orcid.org/0000-0003-0801-0572

(Received 16 March 2026; Revised 18 April 2026, Accepted 30 April 2026; Available online 05 June 2026)

Abstract - **Accurate assessment of student engagement is essential for improving learning outcomes; however, existing engagement analysis methods primarily rely on behavioral logs or self-reported measures, which inadequately capture emotional and contextual dynamics present in real classroom environments. The absence of well-annotated, emotion-aware datasets further limits the development of reliable engagement prediction systems. To address this gap, this study proposes a structured multimodal dataset for student engagement analysis by integrating spontaneous facial emotion data from the Indian Spontaneous Expression Database (ISED) with real classroom video recordings. The dataset is constructed using a systematic methodology involving classroom video acquisition, frame extraction, emotion labelling, and intensity scoring. It comprises 120 participants recorded across 45 classroom sessions, resulting in 320 video samples, approximately 185,000 frames, and a total duration of 8 hours. Each sample is annotated with emotion categories and continuous intensity values, enabling fine-grained affective analysis. Dataset preprocessing and organization are performed using Python, with OpenCV for video processing and NumPy and Pandas for data management. To validate the usability of the dataset, baseline engagement prediction models are implemented using TensorFlow/Keras. Experimental evaluation demonstrates that models trained on the proposed multimodal dataset achieve an accuracy improvement of approximately 4–6% compared to models trained on unimodal or activity-log-based data. The results indicate that incorporating emotional and visual cues significantly enhances engagement representation and prediction reliability. In conclusion, the proposed dataset provides a comprehensive and ecologically valid resource for emotion-aware student engagement research. By offering explicit dataset-level statistics and validated baseline performance, this work supports future advancements in learning analytics, intelligent tutoring systems, and personalized education technologies.**

Keywords: **Classroom Video Data, Dataset Construction, Emotion Score, Emotion Intensity, Student Engagement**

I. INTRODUCTION

Student participation is central in improving academic achievements and the achievement of effective learning. It includes behavioural, emotional, and cognitive engagement in the learning process and affects the manner in which students approach the learning content, peers, and educators (Brahim, 2022). It has always been proven through research that keen students are more likely to enhance their comprehension, improve their levels of retention, and improve their levels of critical thinking (Mazumder et al., 2024; Qarbal et al., 2025). Engagement is a measure of immediate interest in one of the educational processes, but also it is a reliable predictor of future academic achievement and individual growth (Ruiz et al., 2022). Especially in contemporary learning environments, characterised by the use of digital learning environments, measuring engagement and understanding the same may assist teaching professionals in adjusting the teaching content, delivering timely interventions, and supporting a welcoming and encouraging learning environment (Sharma et al., 2022). Although it is of crucial importance, the capability to objectively and in an automatic forecast student engagement is still limited by the presence of strong datasets (Maqsood et al., 2022). Although

there are a number of datasets that can be used to perform facial emotion recognition or behaviour detection alone, significantly fewer publicly accessible, multimodal datasets have been specifically prepared toward student engagement prediction (Feng et al., 2022). Most of the available data is too precise (addressing only facial expressions or individual classroom behaviours) or too deprived of contextual data indicating real attention and engagement of students (Marquez-Carpintero et al., 2025; Mehta et al., 2022). Moreover, not much data allows encapsulating visual (facial and behavioural) and temporal properties with the appropriate level of detail to model engagement correctly (Rizwan et al., 2025). Such shortness invalidates the creation and comparison of AI models capable of generalisation in different classroom environments (Pabba & Kumar, 2024). To fill this gap, it is urgently required to create well-marked and wide-ranging plans by assembling facial expressions, body language, and contextual hints to be able to form the more viable and apposite detection engagement structures in educational settings (Tao et al., 2022; Alhothali et al., 2022).

Among the changes in the realm of digital education, the issue of student engagement is very significant but poorly studied (Xie et al., 2025). Classical methods of detecting engagement mostly depend on predetermined labels of the classes or costly and laborious sensor-based devices, which reduces their scalability and validity in the real classroom case (Thiruthuvanathan & Krishnan, 2025). Also, publicly available multimodal data depicting real-time behaviours of students and emotion expressions in classrooms is in great shortage (Johar et al., 2023; Hossen & Uddin, 2025). Rather than defining engagement in hard and fast categories, the study uses the emotion types and intensities to give a better-grained view of the indication of engagements (Mutharasi & Vijayalakshmi, 2024; Matz et al., 2023).

This work does not concentrate on the construction of deep learning models, but it rather highlights the importance of strict data collection, annotation and labeling processes in providing an excellent base of emotion-aware educational technologies. To cope with the scarcity of multimodal engagement data, the study produces a rich, annotated dataset consisting of spontaneous facial expression data of the Indian Spontaneous Expression Database (ISED) and real-world classroom video recordings (Pabba & Kumar, 2022). It aims to measure in an ecologically robust and structured way the emotional, cognitive and behavioral aspects of student engagement. Unlike the currently available datasets that separate facial expression or classroom behavior, this attempt is a combination of controlled spontaneous emotion samples and spontaneous classroom interaction that would provide a more holistic interaction. The annotation of continuous intensity scores includes emotive analogies, such as happiness, sadness and surprise, and thus affective analysis is more precise when using continuous scores, not categorical. The classroom recordings contain such behavioral cues as gaze, posture, note taking, and work with learning resources. It employs a systematic pipeline, which comprises of multi-angle video capture, peak-frame capture, frame cleaning, frame segmentation and consensus-based annotation of type and degree of emotion. The principles of scaling that have been checked are applied in the conversion of the qualitative affective signals into the quantitative ones to be examined as emotion scores. The resulting data is comprising of heterogeneous participants, multi-session data, organized meta-data and open data on the annotation process. This study fills a gap that is extremely significant within educational analytics due to the quality of the data, emotional granularity, and contextual realism. The main contribution is that it provides a scalable, reusable and carefully edited multimodal data that will support future research studies in the field of learning analytics, affective computing, intelligent tutoring and personalized education. The key contribution of the work is as follows:

- Developed a curated, multimodal dataset integrating ISED and real classroom video data.
- Annotated each sample with emotion labels and computed intensity scores instead of predefined engagement classes.
- Designed a structured dataset partitioned into training, validation, and testing sets for future research use.
- Established a baseline for emotion-based engagement modeling without implementing machine learning models.

1.1 Motivation of the Study

The existing literature on student engagement is mostly based on virtual learning logs, self-reported surveys, or coarse behavioral features, which do not accurately represent the real-time emotional and contextual aspects of learning in the classroom. While some multimodal methods have been proposed, they are mostly constrained by small-scale datasets, controlled settings, and categorical engagement measures without intensity scores for emotions. Many existing methods also focus on model performance rather than the construction and quality of the dataset. This indicates that there is a definite need for ecologically valid and emotion-aware datasets for the classroom setting. This need has led to the development of the proposed dataset, which combines spontaneous facial emotion information from ISED with real classroom video data, annotated with emotion types and continuous intensity scores.

1.2 Rest of Section

- **Section 1: Introduction** – Explains the importance of student engagement and the need for better data to study it.
- **Section 2: Literature Review** – Explains about existing research on student engagement and emotion-based learning systems.
- **Section 3: Proposed Framework** – Describes how the ISED and classroom video data were collected, labeled, and used.

- **Section 4: Dataset Results and Analysis –** Describes about the results and their effectiveness on study
- **Section 5: Conclusion** – Summarizes the study and its value for future educational tools and research.

II. LITERATURE REVIEW

Prior studies in learning analytics have leveraged virtual learning environment (VLE) data to analyze student engagement and academic performance. The Student Engagement Dataset (SED), derived from Moodle-based platforms, aggregates large-scale behavioral data such as login frequency, course enrollment, assignment activity, and temporal usage patterns. Although it will be possible to conduct a statistical analysis of online interactions using such datasets, they will mostly include the records of interactions and academic performance, which can provide little to no information about the affective and emotional feelings of students. Activity measures imply engagement in an indirect way and does not indicate real-time involvement, cognitively or emotionally. This drawback indicates the necessity of datasets that combine behavioural information with emotional and visual features, which encourages the creation of multimodal datasets of emotion-sensitive engagement.

The current literature is based on information gathered on Virtual Learning Environments (VLEs) where the engagement is deduced through activity logs and academic history. These works tend to implement several machine learning classification algorithms in high-level data preprocessing procedures including normalization, encoding and removing outliers. The assessment of model performance based on such measures as accuracy, precision, recall, and AUC is commonly reported as successful in terms of engagement prediction by tree-based models. Although these methods can be considered effective at predictive performance, they are mainly aimed at comparing algorithms and optimization of the outcomes. Engagement is a predicted label based on the behavioral logs that is represented in the final term that is not explicitly characterized by the emotional or visual stimuli that exist in the real classroom setting. Additionally, these studies rely on existing datasets, which provide minimal information regarding the processes of data collection and annotation. These weaknesses suggest that it is desirable to have carefully designed datasets that would provide affective and situational aspects of engagement, which encourages the creation of multimodal, emotion-aware engagement datasets (Alruwais & Zakariah, 2023).

Current literature on learning analytics has used the data of virtual learning environment to analyse student engagement and academic performance in distance learning environment. Among such attempts is the Student Engagement Dataset (SED), which was built using Moodle-based VLE data of a large population in a university. This data is a compilation of comprehensive behavioral data, such as course enrollment, grades, and a record of online activity logs at each time, which can be statistically analyzed to develop large-scale interaction-based patterns. Correlation-based studies out of such data give an idea of behavioral trends which relate to academic results. Nevertheless, these methods are largely based on the indicators of indirect engagement (e.g. frequency of logins and counts of activities) that fail to reflect the emotional and cognitive conditions of learners in the process of learning. Furthermore, the engagement is modeled at an aggregated level, so it is not possible to model fine-grained and real-time learning dynamics. With this gap in mind, the current research is aimed at creating a multimodal dataset combining emotional manifestations and classroom video data to allow representing the student engagement more comprehensively and realistically (Kassim et al., 2025).

The problem of ensuring engagement of the students in online learning settings, which is especially the inability to identify the emotional and attentional states of the students, was discussed in recent studies. To address this drawback, the computer vision-based techniques have been presented to examine visual cues in the process of learning tasks. Among others, there is one effort, which provides a video dataset of online college students working on mathematical problems on a virtual learning platform, with front-facing cameras recording gestures, head pose, and gaze behavior. The annotation of the data is done at frame-level to differentiate between interested and distracted attentions. Deep learning image classifiers and traditional regression-based models used as baseline models are trained to predict student engagement and included in the learning platform to be evaluated. Although the given work shows that it is possible to detect engagement with the help of vision, engagement is modeled with coarse binary labels and is restricted to certain tasks and environments. Moreover, the emotional intensity and the unplanned facial expressions are not clearly modeled. These limitations drive the rationale of more multimodal datasets that combine the intensity of emotion with video recordings in the classroom to have a more complete analysis of engagement (He et al., 2025).

The previous studies have indicated a positive correlation between the general engagement and academic performance of students in a higher education setting using survey-based data and statistical regression analysis. These kinds of studies are usually based on self-reported indices and other demographic factors to examine the dissimilarity in engagement among academic levels and years of study. Although such methods can give some precious empirical evidence on the engagement performance relationships, they fail to record the real-time behavioral or emotional variations in the process of learning activities. Dimensions of engagement, especially emotional and cognitive engagement are typically quantified indirectly, using questionnaires and not through observable indicators. This has resulted in an increased interest in creating structured datasets that can identify spontaneous expressions of emotion and contextual learning behavior using multimodal data sources to allow a more fine-grained and scalable analysis of engagement (Çali et al., 2024).

Most recent studies point to incremental use of new methodologies of instruction in higher education, such as gamification, artificial intelligence, and data mining, to increase student engagement and facilitate personalized learning patterns. MathE has been created as an online learning platform that allows learning to occur at a self-paced, especially in the learning of mathematics, by offering a structured question bank and an adaptive testing system. The data based on these platforms are usually the answers of students to multiple choices, the choice of topic and their performance history, which allows studying the pattern of learning and the efficiency of platforms. Engagement is, therefore, modeled in a narrow behavioral framework, and not in affectual dimensions. These constraints underline the importance of having datasets with emotional and contextual cues and the data on interaction prompting the creation of multimodal datasets that would offer a much more detailed account of engagement (Azevedo et al., 2024).

Earlier researchers have used machine learning to forecast student performance and engagement on the basis of structured data in interactive learning management systems (Lam et al., 2024). The methods are often based on behavioral characteristics, including the duration of time devoted to tasks, activity records, and progress monitoring, which prove the usefulness of analytic in educational evaluation and one-on-one education (Sabuncuoglu & Sezgin, 2023). However, in the majority of the studies, the focus has been on comparison of models and predictive accuracy and not on the emotional, attentional, and visual cues, which play a huge role in the engagement. Available datasets are frequently short, location-specific and lack a variety in time and context, which hampers extrapolation. In addition, emotion intensity, facial expression and classroom dynamics are hardly clearly modeled. These shortcomings make it clear that more multimodal datasets with behavioral, affective and contextual signals should be made richer and scalable to facilitate more comprehensive engagement analysis (Sashank et al., 2023).

Some of the studies conducted on the prediction of student engagement have considered engagement as a multidimensional construct, wherein engagement entails behavioral, emotional, and cognitive involvement. The vast majority of existing studies concentrate on online learning setting, where the level of engagement is measured based on the data on interaction, i.e. the analysis of activity record, time spent, and participation patterns provided by the students. These works usually define the engagement as a set of predetermined categories, such as not engaged, passively engaged, and actively engaged, and use machine learning to predict the engagement rates, such as Decision Trees, Random Forests, Logistic Regression, and Long Short-Memory networks. Performance of a given model is most often assessed in terms of such accuracy-based measures, where the attention is paid to the most efficient predictive algorithm. Nonetheless, these methods are mainly focused on the demonstration of the performance of the algorithm and the prediction of the outcome, whereas engagement is provided as the final categorical label based on the behavioral data only. The emotional moods and visual signals, which are highly significant in influencing student activity, are not given much attention. Furthermore, there is a general lack of focus on the construction of data set, emotional annotation, and actual classroom behavior. These shortcomings underscore the necessity of properly organized, multimodal engagement data that are emotion sensitive (Yan et al., 2025).

TABLE I REFERENCE SUMMARY WITH DATASET, OBJECTIVE, AND SAMPLE SIZE

Reference	Dataset	Objective	Training & Testing Sample Size	Performance Outcome	Limitation
Alruwais & Zakariah, (2023)	Student Engagement Prediction using VLE Data	Predict engagement using behavioral, emotional, and cognitive features	32,593 observations	CATBoost achieved 92.23% accuracy, 94.40% precision, 100% recall, AUC 0.9624; outperformed AISAR baseline	Dataset limited to VLE logs only, no multimodal data (facial/emotional cues absent)
Kassim et al., (2025)	Student Engagement Dataset (SED)	Analyze VLE activity and academic grades	16,609 students; ~12 million data points	Provided correlation analysis; weak but significant relations (e.g., logins vs performance)	Dataset restricted to one semester; no emotional/visual modalities
Çali, et al., (2024)	Multimodal Engagement Dataset	Use facial, keyboard, and mouse inputs to predict engagement	Not specified	Multimodal model achieved higher accuracy and lower MSE vs unimodal	Limited modality scope (no classroom video/audio); synthetic lab setting

Azevedo et al., (2024)	MathE Platform Dataset	Study engagement via multiple-choice answers on a math platform	372 students; 9546 responses	Gamified engagement analysis; allowed personalized learning insights	Only MCQ-based; no multimodal/affective cues
Lam et al., (2024)	VnCodelab Performance Prediction Study	Predict performance via engagement and progress features	253 students	Stacking Classifier outperformed others; highest accuracy among compared models	Small sample size; focused on coding labs only
Sabuncuoglu & Sezgin, (2023)	Human-Centered Engagement Dataset	Analyze audiovisual data with self-evaluation scores	8 hours video segmented into clips	Face-based models**:** 45–85% accuracy; group activity models: up to 71%	Small-scale (only 8 hours); limited classroom diversity
Sashank et al., (2023)	Intelligent Engagement Prediction System	Classify engagement level (not/passive/active) via ML	100 records	LSTM achieved higher prediction accuracy than DT, RF, LR	Very small dataset (100 samples); lacks generalizability
Yan et al., (2025)	Multimodal Engagement Assessment Framework	Fuse multimodal data to assess engagement in blended learning	205 samples (23 high, 147 mod., 35 low)	Deep CNN-based framework achieved high classification accuracy	Class imbalance (majority moderate); small dataset

In table I, the comparative analysis of the student engagement dataset reveals diverse approaches and data sources used to predict or assess engagement. These datasets vary in size and form, including behaviour and educational records to multimodal signals, such as facial expressions, mouse movements and keystrokes (Keinert et al., 2025). Although there are datasets such as (Kasim et al., 2025; Alruwais & Zakariah, 2023) utilize mass VLE activity logs. The sample size is over 12 million pieces of data (to over 100 records) representing a broad range of research parameters and goals that would provide greater learning analytics and adaptive education systems.

III. PROPOSED FRAMEWORK

The framework that is presented in the current research is concerned with the precision and organization of multimodal data that is particularly aimed at guiding the prediction of the engagement degree by the students with the assistance of the emotional analysis. This framework is a synthesis of two significant sources of data, which are the Indian Spontaneous Expression Database (ISED) and a self-provided classroom video collection. The ISED offers high-quality spontaneous facial expression labels for annotation on emotion tags and importance scores to form an effective basis of training. Comparatively, the student video dataset presents classroom interactions found in real life and works as the input in validation, mimicking real learning conditions. The framework entails isolation of peak frames of videos by manually labelling them on emotion type and intensity and then calculating emotion scores using a standardised logic of scale. This emotion-based approach does not use determinate classes of engagement but focuses more on representing the emotion in detail. The dataset will be organised to have training, validation and testing sets and therefore will be consistent and scalable. This architecture creates the foundation for further use in the tasks of the machine learning and intelligent educational systems orientated at the model of emotion-aware engagement. Fig. 1 shows the proposed Framework.

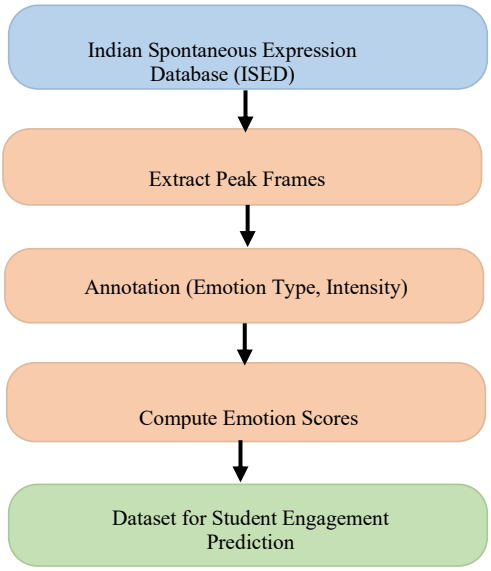

Fig. 1 Proposed Framework

3.1 Dataset Description

In this section, research describe two primary datasets used for the creation of an emotion-labelled resource aimed at supporting the prediction tasks of student engagement. The first dataset is the Indian Spontaneous Expression Database

(ISED), including high-quality video samples of spontaneous facial expressions representing a series of basic feelings such as happiness, sadness, anger, surprise and neutrality (Happy et al., 2015). It acts as a fundamental source to learn dataset emotion-related features. In the second dataset, there is a classroom video recording collected from real student interaction during learning sessions, which captures natural behaviour and emotional manifestations. These recordings were achieved with diverse participants and camera angles and with different engagement levels, yet under realistic conditions. Together, these datasets provide a comprehensive mixture of controlled emotional signs and real-world behaviour, enabling the development and verification of intensity-based engagement analysis of emotions.

Classroom Student Video Dataset is the result of an organized experimental design, the purpose of which was to observe natural behavioural and emotional expressions to indicate different degrees of involvement in classroom tasks. To ensure the ecological validity and also to ensure that there was control in the experiment, the students were provided with specific learning and collaborative activities and the data were recorded under two different conditions, i.e., Session I (video-based activity) and Session II (tool-based activity). During the first session (Session I), the students were exposed to instructor-led concept-based discussions and it was noted that they were attentive in terms of facial expression, gaze direction, posture and gestures as they followed brief educational video modules. The camera was strategically positioned in three fixed angles front-facing camera, which would capture micro-expressions and direct attention, a side view camera, which was used to observe the dynamics of gestures and partial profile cues, and a top-corner camera to capture the interactions between groups in general and their engagement. These multi-angle placements belong to the established practices of engagement-oriented video to guarantee the complete coverage of space and align with the previous research on multimodal learning analytics.

During the second session (Session II), students completed the tasks of problems solving using tools, and the interactions of students with digital tools, notes, and shared resources were filmed. This session was to produce various responses affective and behavioural as opposed to passive video observation. The combination of the two sessions, in which one is based on perceptual attention and the other is an interactive interaction, results in a balanced set of data, which includes cognitive, emotional, and behavioural aspects of engagement. All of them were recorded with high-definition (1280472 pixels, 30 FPS) cameras under controlled classroom lighting so that facial clarity was not lost. The cumulative time spent during all the recording sessions went beyond eight hours. Videos were subsequently collected and the videos reviewed manually to isolate single frames where facial areas were clearly visible. The Occlusion or moving blurs frames were discarded. The segments were grouped into clips of 10 seconds and manually annotated by trained raters who were given a score on the type of emotion (e.g., happiness, surprise, sadness, and disgust, neutral) as well as the intensity of the emotion, based on consensus, to achieve inter-rater reliability. The well-designed data-collection design allows a smooth correspondence between the controlled data on emotion provided by ISED and spontaneous data on behaviour in the classroom set-ups. The availability of multi-angle recordings, predetermined activity guidelines, and dual-session engagement situations guarantees the appropriateness of the dataset to a multimodal analysis in the future and the creation of emotion-sensitive engagement prediction algorithms. Subsequent to this, this dataset is the empirical basis to the actualization of the suggested framework presented in later sections to support the advancement of the raw observation to intelligent feedback creation in the educational AI systems.

3.1.1 Indian Spontaneous Expression Database (ISED)

The Indian Spontaneous Expression Database (ISED) consists of two major parts – video and peak images – both of which are aimed at capturing and presenting spontaneous emotional expressions by the participants in a defined and descriptive form. The central part of the dataset is the video recordings. Every participant makes some short video sequences at hand, and the latter demonstrate the natural unfolding of one in a row of emotional facial interactions. These expressions build up to a neutral baseline, then to emotional intensity; they represent the entire time events that happen on the face. Such development is essential to training models on the dynamics of emotions changing through time and not just on the recognition of unchanging changes in the face. Video data is also organised in a systematic way so that all participants have a dedicated folder under the videos/ directory named PS. Each folder contains individual video clips, which are named in sequence (e.g., 4.avi is the fourth video clip). An example is the file videos/PS 1/4.avi, which shows that this is the fourth video of recording emotional expressions in Participant 1.

Every video is annotated with one of four basic emotion classes:

- 1: Happiness
- 2: Surprise
- 3: Sadness
- 4: Disgust

The labels are given according to the spontaneous emotional response recorded in each video and are identical throughout all files included in the dataset, which makes it convenient to map and refer to. The ISED also has photos of the peak (static frames that signify the point of strongest emotional impact in each video sequence). The choice of these frames and extraction is made by careful visual inspection or defined thresholds on the intensity. Being saved in the zip archive called peakImages.zip, these images provide a concise but very informative perspective of the emotion that is presented in the video. The vocabulary used in the names of the images of peaks is dry and logical: each image has the following name: Participant_Clip_Frame.jpg. For example, file

26_7_39.jpg participants corresponds to 26, the 7th video clip and the 39th frame of that clip, which has been identified as the point of peak emotional expression. The researchers do them using dual-formal-structure-temporal video and high-spuriousness, still frame-flexibility. The dynamics of temporary expression can be analysed using video data and snapshot-based classification or training image-level recognition models can be done using the peak paintings. The dataset provided by this organisation is very appropriate to be used in multimodal analysis, and it can be generalized in different tasks in machine learning, including supervised classification, feature extraction, and the training of deep learning models.

The .mat file named ISED_details.mat dataset, the metadata is given a structured form to allow analysis and programmatic access. This file contains tabular information containing the following information about each sample:

TABLE II OVERVIEW OF METADATA FILE

Column	Description
1	Participant Number (e.g., 1, 2, ..., 50)
2	Unique Participant ID
3	Video Number
4	Peak Frame Number in the video clip (starting from frame 1)
5	Full Video File Path
6	Full Peak Image Path
7	Emotion Label (encoded as: 1 = Happiness, 2 = Surprise, 3 = Sadness, 4 = Disgust)
8+	Facial Landmark Coordinates: includes pixel positions of the face bounding box, and key points such as nose, left eye, and right eye in the peak frame

Table II ISED_details.mat, an essential component of the dataset as it includes detailed annotations and structural data and makes the dataset simpler to navigate and analyse in the most productive manner. The various fields in this metadata table describe the individual video clips and their associated peak image. The first represents the number of participants, whereas the second column assigns a unique ID to the participant to make associating data instances quite clear. The third column represents the video number per participant, followed by a peak frame number, the frame in the video sequence currently demonstrating the strongest degree of emotion. The fifth and sixth columns give the full file paths of the original video and the peak image, respectively, associated with this. Column 7 is the label of emotion, since it is coded in a numerical way with 1 representing happiness, 2 surprise, 3 sadness and 4 disgust. Since the eighth column, the metadata contains coordinates of facial landmarks, that is, the pixel coordinates of the face-bounding box, nose, left eye, and right eye in the peak frame. This metadata makes it easy to map video data and visual information, and this is essential in training emotion recognition models, defining datasets, annotating datasets, and analysing at the frame level or feature level.

Fig. 2 Sample ISED Images

Some representative sample images from the Indian ISED are presented in fig. 2, the quality and variety of the spontaneous facial expressions captured in the dataset. These pictures correspond to the frame of emotional intensity extracted from the video clip and reflect a series of feelings such as happiness, surprise, sadness and hatred. Visual examples

highlight the ability of a dataset to catch subtle and natural facial signals in realistic conditions, which are essential for training emotion recognition models. Each image displays the activities and manifestations of the facial muscles associated with different emotional stages that provide a visual reference to the types of characteristics that can be learnt and analysed in the prediction functions of engagement.

When referring to this study, it can be said that the procedure of measuring affective reactions based on the ISED includes the scope of converting crude annotation into a valuable score explained as the Emotion Score. Calculation of this score involves two basic components that are found in the dataset: the emotion label, which identifies the type of emotion viewed, and the emotion intensity, which refers to the degree of how the emotion is portrayed through a particular frame. This transformation would have the goal of normalising emotional clues to one numerical value that represents the level of perceived engagement of the emotion expression. All the termed emotions indicate a broad type of affective response, and rather than boxing them through a categorical set, the present study defines each term in terms of how they may contribute to and represent engagement levels, like low, medium or high. An individual scoring formula is used against each type of emotion to provide an accurate score on the level of the emotion in terms of its normal strength and applicability in real-life learning contexts. As an example, emotions such as happiness are assigned larger values in the base, and that upper value is 100, which shows that those emotions are highly linked to active engagement. On the contrary, less favorable affective traits, including sadness or disgust, are dimensioned in a more conservative way (with lower score ceilings) in order to reconcile with the common view of their connection with disengagement or passive states. This scale technique will help to convert the qualitative data of the emotions to a more quantitative format that is more compatible with machine learning and statistical analysis. Following the numerical scoring of the facial expressions, the research makes it possible to incorporate the emotional information into the engagement prediction models. These scores are also used in reference to the labelling of video clips in the student dataset; this is done to make sure there is consistency across controlled emotion samples (ISED) and spontaneous classroom recordings. The last outcome is that the emotion score will be a useful item in the creation of a robust emotion-aware engagement detector that would serve in an educational setting successfully.

One of the most important functions in this simplification is the compute_emotion_score(row) one that has the capability of converting raw emotion records into a standard numerical value in what is termed Emotion Scores. The dataset contains two significant fields each in a row: Emotion Label Emotion Label indicates the emotion that was identified in a video frame (e.g. happiness, surprise, sadness or disgust). Rather than viewing these as given categorical emotional states, this research interprets these states as relevant to the levels of engagement or in terms of being high, medium or low engagement. The min() function is used in each formula to prevent the resulting score to be larger than its assigned maximum value. This normalization facilitates the consistency of different types of emotions, as well as the differing intensity levels of said emotion, as such this normalization facilitated a smooth integrative process of emotion types and intensities into machine learning models used to predict engagement. After calculating these emotion scores, they can be saved in a separate column Emotion_Score that can further be used as a feature to train predictive models or study some emotional trends in a dataset of classroom engagement.

Particularly, the emotion labels and peak intensity frames of ISED are utilized to:

- Train models to be able to tell different emotional states apart
- Extract facial expression and intensity of emotion features
- Provide a reference model to support emotion-level annotation in the student video dataset
- Provide a baseline for emotion classification before validating engagement predictions on actual student data

3.1.2 Classroom Student Video Dataset

Student Video Dataset was selected to be designed in a way that records natural classroom interactions and emotional cues that indicate engagement in real world learning settings. These recordings are the real dynamics of a classroom unlike the controlled ISED dataset as they provide a good contextual-level of information on emotion-to-engagement mapping. The information was gathered during 8 hours of real classroom sessions of undergraduate students of different gender and academic backgrounds. The videos were captured at 1280×720 at 30 FPS at the consumer cameras (smartphones/webcams) that were staged in various positions:

- Front-facing cameras: recorded overt facial responses and micro-expression.
- Top-corner cameras: offered broad based group interaction and general involvement.
- Side-angle cameras: centered on gestures, direction of gaze, and posture.

This multi-angle design allowed to have both individual facial parameters and group-wide behavioral parameters to analyse. Nevertheless, clean student frames (those in which the face of one student was distinctly visible without occlusion) were annotated only. Frames that had group shots, faces that were blurred or partly obscured features were deleted to ensure that the quality of data was not compromised. All retained clips were divided in 10-second windows and labelled with:

- Label of emotion (e.g., happy, sad, surprised, disgusted, indifferent).
- An intensity score of emotion, which is rated by a variety of annotators to ascertain inter-rater reliability.

This data set thus supplements ISED with ecologically robust classroom data to fill the gap between laboratory controlled and real-life educational settings. The following fig. 2 shows the Classroom Student Video Dataset collection process:

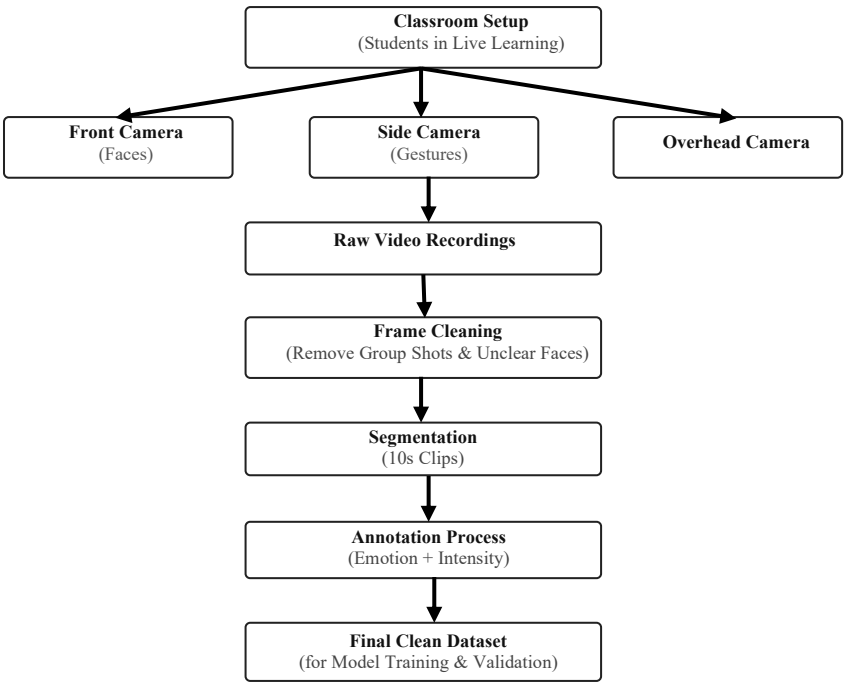

Fig. 3 Classroom Student Video Dataset Collection Process

The collection of the data under the Student Video Dataset in this work was self-collected to aid the real-life modelling of the student engagement by recording the natural classroom physical behaviour as mentioned in fig. 3. The collection was selected during several classroom lessons taking place in a specially controlled academic environment. The video recordings recorded a total of around 8 hours with undergraduate students taking an active role in classroom activities. The group of participants is rather diverse, both in gender and academic status, which offers a wide outlook on classroom activity and emotional involvement. A consumer-grade camera, such as a smartphone camera or webcam, was used to record the video data with a resolution of 1280×720 pixels at 30 frames per second (FPS). The recording system consisted of various camera angles, one with a front view to get direct facial reactions and others at the upper corner positions to gain more of the overall interaction of the groups. Such a multi-angle strategy ensured imaging of individual and group-level behaviour but also created complexity in the analysis of the data in the presence of occlusions and competing facial features when imaging groups. First, raw video data material consisted of a lot of cases of group shots with many students in one frame. They were not considered the frames that could be used to analyse individual emotions and participation because facial features and expressions data isolated were not honest anymore. In order to guarantee the quality of the data and use of individual-level annotation, researcher then manually removed all the footage and frames with group shots or obscured scenes and only kept those scenes in which the facial region of only one student was observed perfectly clearly.

Fig. 4 shows samples of students in the dataset. From clean data, the video clip was cut into manageable segments usually sampled for 10 seconds in length and frame-tier analysis. All these frames were then annotated by hand with two important attributes, an emotion label (e.g. happiness, sadness) and a rating of feeling intensity, which is a score of the level of emotional expression. To improve the validity of the results, several annotators viewed the same segment, so there was a chance to come to a consensus about the labeling of emotions and intensity scores. This collaborative method not only reduced bias but also diversified the information on the emotional expression. It is used as the main verification in self-conscious student video dataset research. It also fills a gap between the controlled, emotion-coloured ISED data and the dynamic, in the field real-world classes, which provides a good basis to test the model models of training and engagement.

Fig. 4 Students Video Sample

3.2 Dataset Structuring

The data on the research are applied to this research and are harmoniously structured to allow the creation and performance analysis of emotion-sensitive engagement forecasting models. It contains the sequences of facial images, small video sequences and manually calculated feature vectors, and the respective emotion labels and intensities accompany them. These annotations are provided on a frame or segment level and patterned representation of them has high resolution in emotional conditions. The data will be divided into two parts one (80 %) to train the model and to validate, and the other (20 %) to test the model, and finally testing (20 %) to assess the performance. In particular, regarding the validation set, real-world data on the student video is used to bring the set into the context of a real classroom. In addition, each data point will be enhanced by metadata (age group of a participant, type of activity in the classroom and the environmental conditions). This syntactic structure can be applied in a predictable manner to model pipelines, and it can be applied to better report on performance within different learning contexts, which facilitates scalable experimentation and informed interpretation regarding student engagement tendencies.

Fig. 5 Emotion Sample

Fig. 5 displays a grid of six photos, each of which has a person's face. The photos are arranged in two rows of three. Each image is labelled with a specific feeling: happiness, sadness or hatred. The top line expresses happiness to a person, then the person expresses grief, and a third person expresses hatred. The bottom line also expresses happiness to a person, another person expresses happiness, and finally, the person expresses sadness. This grid appears to be a visual dataset used for training or testing of a visual recognition system, where each image acts as an example of a particular facial expression that suits a given feeling.

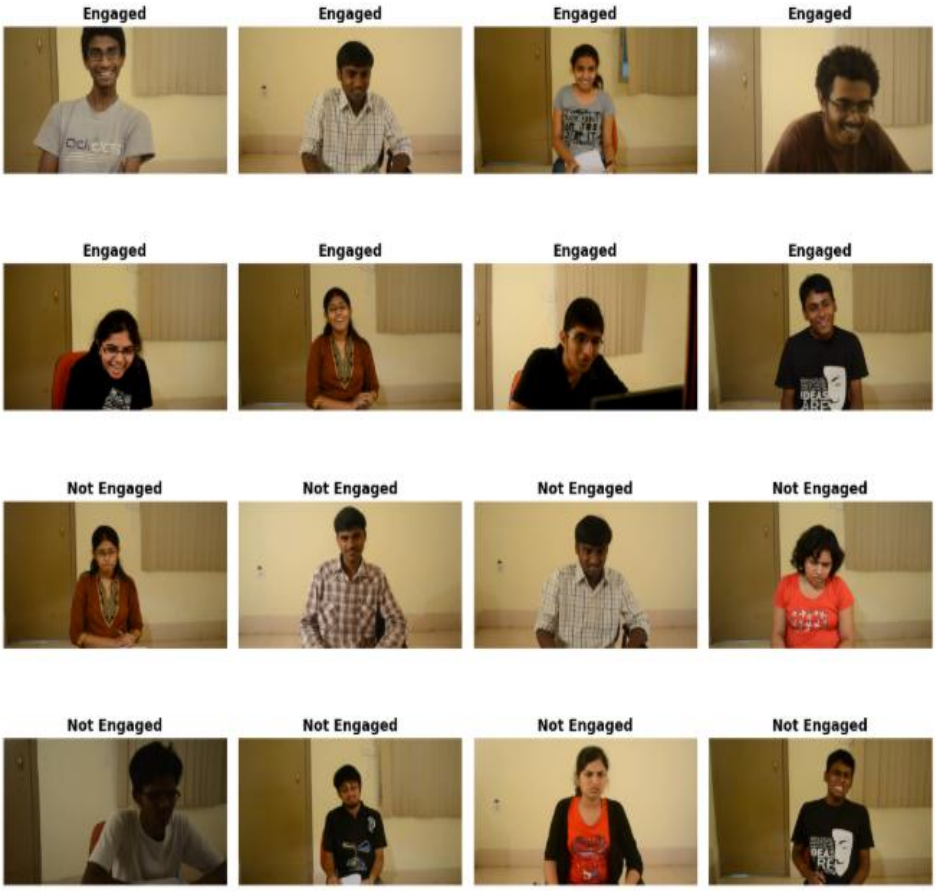

Fig. 6 Student Engagement Sample

Fig. 6 shows a grid of sixteen photos with four rows and four columns; each is labelled to display either a planned or engaged state. The top two rows include eight separate individuals, all of whom are labelled as attached, displaying facial expressions and currencies that suggest that they are attentive or involved. In contrast, the two lines below present eight other individuals, which are not attached, with facial expressions and body language that indicate attention or lack of interest. This layout strongly suggests that the image is a dataset used to evaluate training or machine learning models, designed to detect a person's engagement level, possibly based on visual signs such as facial expressions and body gesture.

In this study, the dataset is strategically divided to adapt to model development and evaluation in various sources as mentioned in table III. The training set contains 80% of the Indian Spontaneous Expression Database (ISED), which provides a rich foundation of the comfortable facial expressions annotated with emotional labels and intensity scores to effectively train the model:

TABLE III DATA SPLITTING

Round	Dataset	Ratio	Description
Round 1	ISED	Training 80% and Testing 20%	Consists solely of 80% of the ISED dataset, used to train the model on spontaneous facial expressions with annotated emotion labels and intensity scores.
Round 2	ISED and Classroom	Training ISED and Testing Classroom	Comprises 20% of the self-collected student video dataset, used exclusively for model validation and hyperparameter tuning in real classroom contexts.
Round 3	Merged ISED and Classroom Data	Training ISED+Classroom 80% ISED+Classroom Testing -20%	The remaining 20% of the ISED dataset, kept unseen during training, used to evaluate the model's generalization performance.

The verification set contains 20% of the self-eccentric student video dataset, which represents the real-world classroom landscapes and serves to fix the hyperparameter, assessing the adaptability of the model for practical learning environments. Finally, the test set contains the remaining 20% of the data issued, which is completely ignored during training and is used to objectively evaluate the normalisation capacity of the model. This division ensures a strong training process, validating the performance of the model under realistic conditions and maintaining fairness in the final assessment.

3.3 Annotation and Labelling Strategy

The process of annotation and labelling in this research is aimed at recording the level of emotions and their degrees of intensity but not the level of engagement itself. Via this method, more multi-dimensional data is produced, which can be used in a range of emotion-aware prediction of engagement applications. The main annotation process objective lies in labelling each of the frames or segments of the video that will be contained in the dataset with two important attributes:

- Emotion Label – the emotion that is expressed in the frame (e.g., happy, sad, angry, neutral, surprised, or disgusted).
- Emotion Intensity Score – this is the numerical rating of the strength or vividness of the emotion presented.
- Frame and Segment-Level Marking

Annotations took place at the frame level or short video segment level, according to the presence of clarity and consistency in emotional display. There was a frame-level labelling that was employed in instances where the emotions were changing at a high rate. In segments that contained the same emotional state, the same annotation was used on all the frames in such a segment. Granularity was chosen in a way that it provided accuracy in labelling and efficiency in the annotation process.

The annotation of data was done using well-trained human annotators who are already exposed to the concept of facial expression recognition. A set of standardised guidelines, including some examples of each and every emotion category, was explained to the annotators to diminish the subjective variability.

- Step 1: Observe the video frame or segment to see direct facial expressions of eyebrow movement, eye opening, mouth shape and the head direction.
- Step 2: Determine the emotion that is mostly revealed. In case of ambiguity of the expression, the annotators referred to the most salient visual indication.
- Step 3: Allocate an intensification score depending on how strong an emotion is perceived to be. Bear in mind that this decision is influenced by the tension shown in the facial muscles, how expressive someone was and how long or how short the emotion lasted.
- Step 4: Write down the label as well as the intensity score in the annotation sheet of the data set.

Three annotators to independently reduce bias annotated each frame or section. For each example, the last label was determined using a consensus-based approach:

- If all the annotators agreed, the label was directly accepted.
- If there was disagreement, the majority vote set the label.
- In matters of equally divided opinions, a discussion was held to reach the agreement, and the frame was re-examined for clarity.

Final feeling intensity score of every frame was obtained as mean value of the scores of three annotators which were a balanced remedy, and which brought about the variance in personal perception.

IV. Dataset Results and Analysis

This section will contain a more detailed analysis of the proposed multimodal dataset of student engagement in its statistical properties, annotation features, and patterns identified. The data obtained is displayed at the level of the datasets to provide the clarity of the composition and quality of the obtained data. The category of emotion frequency and intensity of variation of sample distribution are regarded as the important points to highlight the representativeness of the dataset. It also establishes the consistency and reliability of annotation in order to determine the strength of the labelling process. The contextual and temporal trends of the sessions in the classroom are also provided.

	Participant Number	Participant ID	Video Number	Peak Frame Number	Video File Path	Peak Image Path	Emotion Label	Emotion Intensity
0	51	'PS_51'	11	142	'./videos/PS 51/11.avi'	'peakImages/51_11_142.jpg'	1	3.75
1	9	'PS_9'	13	87	'./videos/PS 9/13.avi'	'peakImages/9_13_87.jpg'	1	4.75
2	24	'PS_24'	6	43	'./videos/PS 24/6.avi'	'peakImages/24_6_43.jpg'	1	3.75
3	3	'PS_3'	16	21	'./videos/PS 3/16.avi'	'peakImages/3_16_21.jpg'	4	4.25
4	45	'PS_45'	4	215	'./videos/PS 45/4.avi'	'peakImages/45_4_215.jpg'	4	3.25

Fig. 7 Sample Participant Detail

Fig. 7 above the Indian spontaneous expression database (ISED) presents a sample view of the metadata, describing the key features for each recorded example. Each row corresponds to a unique emotional expression captured from a participant, identified by their participant number and ID. Metadata contains the video clip number and peak frame number, which indicates the exact frame where it reaches its highest intensity. Full video (Vi) and extracted peak image (.JPG) file paths are also made readily accessible and accessible. All of the examples are labeled with an emotion category, which identifies the categories of expressions made, and a constant emotion on a similar scale is a similar emotion intensity score. As an illustration, the first-row participant presents the 11th video clip of 51, in which the best expression can be found on the frame 142, marked with a 3.75 feeling. This Structured Format is helpful in terms of exact analysis and effective incorporation into machine learning pipes.

TABLE IV EMOTION SCORE CALCULATION

Emotion Label	Associated Engagement Level	Scoring Formula	Maximum Score Cap
Label 1	High	Min (100, 80 + intensity × 4)	100
Label 2	Medium-High	Min (80, 60 + intensity × 4)	80
Label 3	Medium	Min (60, 40 + intensity × 4)	60
Label 4	Low	Min (35, intensity × 8)	35

Emotion Intensity: table IV shows the degree of the expression of the emotion in the specific frame and it usually varies between no value and the maximum value dependent on the annotations in particular data. Individual mathematical formulas are implemented to each type of emotion to use in calculating the Emotion Score. These formulas normalize the strength of each emotion to a limited number that displays its average influence on the engagement of the students in academic settings.

For instance:

- Label 1, by extension to high engagement, has a high base value that is limited to 100.
- Label 2 is also associated with a somewhat positive engagement cue, but with a bit lower cap.
- Label 3 is instead a moderate indicator and hence possesses a lower range of scores.
- Label 4 which is highly correlated with disengagement or negative reaction is rated towards the lower end.

In this study, instead of relying on pre-degraded engagement categories, each sample is annotated with two major indicators: emotion type – like happiness, sadness, surprise, or disgust – and a feeling intensity score, which is a constant value received from human annotations and a custom scaling formula. This dual labelling method gives a more detailed and adaptable depiction of the emotional stages, and data is able to record the minor differences in expression and involvement in emotion. This fine-stable labelling allows more precise and adjustable modelling since such broad emotional indicators will be later mapped at a broader level of engagement depending on particular application needs or experimental design.

The input formats in this study are intended to be effective in representing and storing the dynamic and fine nature of student emotions. These are sequences of facial images, which are obtained out of the ISED and classroom video data to record the temporal change in facial expression. Every image is normalised and standardised (e.g. in .JPG or .PNG format) so that it is stable through model training. Moreover, video clips are typically employed to maintain the natural sequence of emotional transitions, are approximately 10 seconds long, and contain some metadata like resolution (1280x720), frame count, and participants. To facilitate quantitative analysis, feature vectors are extracted on peak-frames and image-series, such as feature landmark coordinates, gaze direction as well as manual annotated intensity values. These vectors are represented in structured files like .CSV or. NPY thus allowing easy incorporation in machine learning pipelines in order to forecast emotion-incinerated engagement.

Fig. 8 depicts the chosen sample record out of the end structured dataset that illustrates how the raw emotion labels and intensity have been transformed into the standardised emotion scores that can be used downstream. Each row is a specific participant and video clip, with the fields of participant numbers and metadata, such as ID, video number and peak frames, where the emotional expression is most significant. Video and image file path models Model of related video and image file path are used to refer to the related media resources to be used in training and verification. The emotional label identified identifies the type of emotion and the intensity of feelings is expressed as the intensity of emotions in a scale. The notable findings in the figure is the emotion-skor that is determined by the custom scaling logic in relation to each type of emotion. For example, the participant expresses a feeling with a label of 421 and an intensity of 2.75, resulting in a high engagement score of 91.0. This integrated visual ensures a consistent and quantitative representation of emotional states, enabling more accurate modelling of student engagement.

	Participant Number	Participant ID	Video Number	Peak Frame Number	Video File Path	Peak Image Path	Emotion Label	Emotion Intensity	Emotion_Score
205	9	'PS_9'	3	22	'/videos/PS 9/3.avi'	peakImages/9_3_22.jpg	4	4.75	35.0
288	11	'PS-11'	18	163	'/videos/PS 11/18.avi'	peakImages/11_18_163.jpg	3	3.50	54.0
380	42	'PS_42'	4	119	'/videos/PS 42/4.avi'	peakImages/42_4_119.jpg	1	2.75	91.0
206	3	'PS_3'	9	28	'/videos/PS 3/9.avi'	peakImages/3_9_28.jpg	2	4.00	76.0
338	35	'PS_35'	12	135	'/videos/PS 35/12.avi'	peakImages/35_12_135.jpg	4	2.75	22.0
68	24	'PS_24'	5	41	'/videos/PS 24/5.avi'	peakImages/24_5_41.jpg	1	4.00	96.0
305	45	'PS_45'	11	126	'/videos/PS 45/11.avi'	peakImages/45_11_126.jpg	1	4.25	97.0
162	4	'PS_4'	9	153	'/videos/PS 4/9.avi'	peakImages/4_9_153.jpg	2	3.50	74.0
142	49	'PS_49'	3	60	'/videos/PS 49/3.avi'	peakImages/49_3_60.jpg	1	3.25	93.0
88	30	'PS_30'	6	201	'/videos/PS 30/6.avi'	peakImages/30_6_201.jpg	1	4.00	96.0

Fig. 8 Sample Records with Computed Emotion Scores

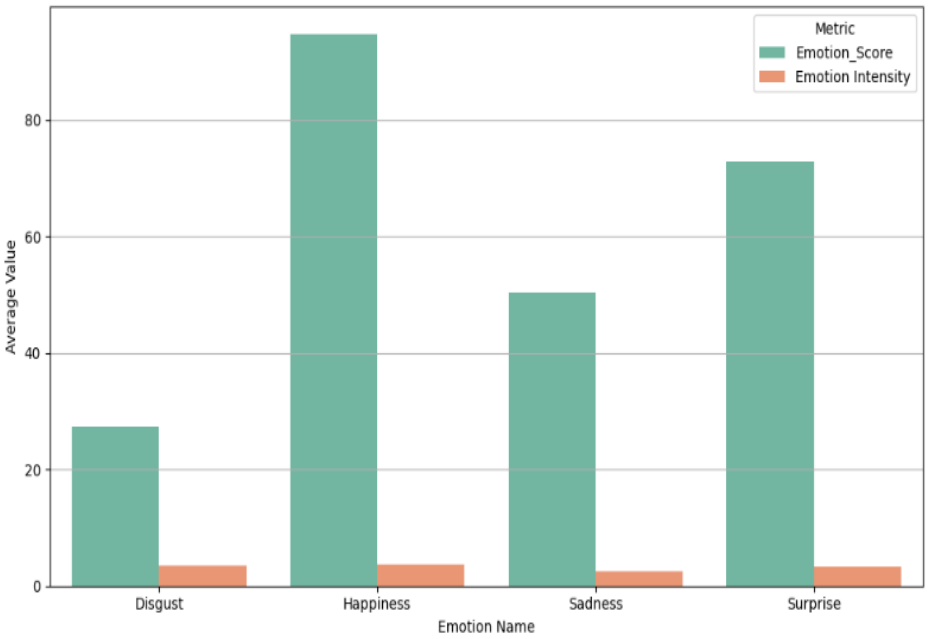

Fig. 9 Average Emotion Score Vs. Intensity by Emotion

Fig. 9 presents a comparison between the average emotion score and emotion intensity for four different emotions: disgust, happiness, sadness and surprise. For each feeling, there are two. Teal-coloured bars represent the average emotion score, which is much more consecutive than small, coral-coloured bars representing the average emotion intensity. Happiness has the highest emotion score, then surprise, then sadness and finally disgust. In particular, the emotion intensity score for all four emotions is very low and looks similar to each other, showing that while the model or system is confident in its emotional scoring, the intensity of those feelings in data is generally weak.

V. CONCLUSION

A hierarchical multimodal database to enable emotion-conscious student interaction evaluation in the classroom learning setting. The dataset has 120 participant records, 45 classroom sessions, which cover 320 video samples and 185,000 frames totalling to 8 hours. Well-defined categories of emotions and continuous scores of intensities are annotated on each sample giving the ability of fine-grained representation of the engagement dynamics that go beyond the coarse categorical labels. Analysis of data-sets showed realistic distribution patterns of emotions, natural imbalance of classes, and regular annotation patterns,

all of which was a representation of natural behavior in a classroom. Temporal observations also showed the change in emotional intensity per learning session, which is dynamic in terms of the student engagement. In addition to numerical characterization, the key value addition of this work is that it creates a reusable and extensible research resource of the learning analytics community. As opposed to the current datasets that are based on the evaluation of behavioural records only or self-reported scales, the offered dataset is a combination of visual, emotional, and contextual data, making it more ecologically valid and analytical. The dataset will be applicable to a general deployment of diverse future applications, such as engagement prediction, affective computing, intelligent tutoring systems, and personalized learning analytics. Using accessible statistics of dataset and standardized annotation, with baseline validation, the work sets a strong foundation of reproducible and scalable research in student engagement to fill a significant gap in emotion-aware data on education.

Future Works will be improved by adding more participants to the dataset, both classroom sessions and subject diversity to increase the generalizability. Audio cues and physiological cues can also be introduced to increase engagement representation through other modalities. It is possible to use the high-level deep learning and temporal modeling techniques to compare the engagement prediction and help it with the assistance of the dataset. Also, the publishing of standardized evaluation protocols will help in just comparisons and reproducibility in subsequent studies on engagement.

REFERENCES

[1] Alhothali, A., Albsisi, M., Assalahi, H., & Aldosemani, T. (2022). Predicting student outcomes in online courses using machine learning techniques: A review. *Sustainability*, *14*(10), 1-23. https://doi.org/10.3390/su14106199

[2] Alruwais, N., & Zakariah, M. (2023). Student-engagement detection in classroom using machine learning algorithm. *Electronics*, *12*(3), 731. https://doi.org/10.3390/electronics12030731

[3] Azevedo, B. F., Pacheco, M. F., Fernandes, F. P., & Pereira, A. I. (2024). Dataset of mathematics learning and assessment of higher education students using the MathE platform. *Data in Brief*, *53*, 110236. https://doi.org/10.1016/j.dib.2024.110236

[4] Brahim, G. B. (2022). Predicting student performance from online engagement activities using novel statistical features. *Arabian Journal for Science and Engineering*, *47*(8), 10225-10243. https://doi.org/10.1007/s13369-021-06548-w

[5] Çali, M., Lazimi, L., & Ippoliti, B. M. L. (2024). Relationship between student engagement and academic performance. *International Journal of Evaluation and Research in Education*, *13*(4), 2210-2217. https://doi.org/ 10.11591/ijere.v13i4.28710

[6] Feng, G., Fan, M., & Chen, Y. (2022). Analysis and prediction of students' academic performance based on educational data mining. *IEEE Access*, *10*, 19558-19571. https://doi.org/10.1109/ACCESS.2022.3151652

[7] Happy, S. L., Patnaik, P., Routray, A., & Guha, R. (2015). The Indian spontaneous expression database for emotion recognition. *IEEE Transactions on Affective Computing*, *8*(1), 131-142. https://doi.org/10.1109/TAFFC.2015.2498174

[8] He, Y., Feng, Z., & Liu, W. (2025). A Multimodal Analysis of Automotive Video Communication Effectiveness: The Impact of Visual Emotion, Spatiotemporal Cues, and Title Sentiment. *Electronics*, *14*(21), 4200. https://doi.org/10.3390/electronics14214200

[9] Hossen, M. K., & Uddin, M. S. (2025, February). A Multimodal Monitoring System with XGBoost Classifier for Optimizing Student Engagement in Online Learning. In *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (pp. 1-6). IEEE. https://doi.org/10.1109/ECCE64574.2025.11012991

[10] Johar, N. A., Kew, S. N., Tasir, Z., & Koh, E. (2023). Learning analytics on student engagement to enhance students' learning performance: A systematic review. *Sustainability*, *15*(10), 7849. https://doi.org/10.3390/su15107849

[11] Kassim, M. S. S., Azizul, Z. H., & Fuaad, A. A. (2025). Student engagement dataset (SED): An online learning activity dataset. *IEEE Access*, *13*, 23607-23617. https://doi.org/10.1109/ACCESS.2025.3531102

[12] Keinert, M., Pistrosch, S., Mallol-Ragolta, A., Schuller, B. W., & Berking, M. (2025). Facial emotion recognition of 16 distinct emotions from smartphone videos: comparative study of machine learning and human performance. *Journal of Medical Internet Research*, *27*, e68942. https://doi.org/10.2196/68942

[13] Lam, P. X., Mai, P. Q. H., Nguyen, Q. H., Pham, T., Nguyen, T. H. H., & Nguyen, T. H. (2024). Enhancing educational evaluation through predictive student assessment modeling. *Computers and Education: Artificial Intelligence*, *6*, 100244. https://doi.org/10.1016/j.caeai.2024.100244

[14] Maqsood, R., Ceravolo, P., Romero, C., & Ventura, S. (2022). Modeling and predicting students' engagement behaviors using mixture Markov models. *Knowledge and Information Systems*, *64*(5), 1349-1384. https://doi.org/10.1007/s10115-022-01674-9

[15] Marquez-Carpintero, L., Viejo, D., & Cazorla, M. (2025). Enhancing engineering and STEM education with vision and multimodal large language models to predict student attention. *IEEE Access*. *13*, 114681–114695. https://doi.org/10.1109/ACCESS.2025.3584025

[16] Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., Dinu, A., & Stachl, C. (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Scientific Reports*, *13*(1), 5705. https://doi.org/10.1038/s41598-023-32484-w

[17] Mazumder, D., Chatterjee, A., Chakraborty, A., & Karmakar, R. (2024, June). A Novel Student Engagement Level Detection and Emotion Analysis Using Ensemble Learning. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-7). IEEE. https://doi.org/10.1109/ICCCNT61001.2024.10725730

[18] Mehta, N. K., Prasad, S. S., Saurav, S., Saini, R., & Singh, S. (2022). Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement. *Applied Intelligence*, *52*(12), 13803-13823. https://doi.org/10.1007/s10489-022-03200-4

[19] Mutharasi, T., & Vijayalakshmi, S. (2024, December). A Multimodal Deep Learning Approach to Decoding Student Participation in Online Classrooms. In *2024 OITS International Conference on Information Technology (OCIT)* (pp. 149-154). IEEE. https://doi.org/10.1109/OCIT65031.2024.00035

[20] Pabba, C., & Kumar, P. (2022). An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. *Expert Systems*, *39*(1), 1-28. https://doi.org/10.1111/exsy.12839

[21] Pabba, C., & Kumar, P. (2024). A vision-based multi-cues approach for individual students' and overall class engagement monitoring in smart classroom environments. *Multimedia Tools and Applications*, *83*(17), 52621-52652. https://doi.org/10.1007/s11042-023-17533-w

[22] Qarbal, I., Sael, N., & Ouahabi, S. (2025). Student's engagement detection based on computer vision: A Systematic Literature Review. *IEEE Access*, *13*, 140519–140545. https://doi.org/10.1109/ACCESS.2025.3596885

[23] Rizwan, S., Nee, C. K., & Garfan, S. (2025). Identifying the factors affecting student academic performance and engagement prediction in mooc using deep learning: A systematic literature review. *IEEE Access*, *13*, 18952-18982. https://doi.org/10.1109/ACCESS.2025.3533915

[24] Ruiz, N., Yu, H., Allessio, D. A., Jalal, M., Joshi, A., Murray, T., ... & Betke, M. (2022). ATL-BP: a student engagement dataset and model for affect transfer learning for behavior prediction. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *5*(3), 411-424. https://doi.org/10.1109/TBIOM.2022.3210479

[25] Sabuncuoglu, A., & Sezgin, T. M. (2023). Multimodal Group Activity Dataset for Classroom Engagement Level Prediction. https://doi.org/10.48550/arXiv.2304.08901

[26] Sashank, Y. T., Kakulapati, V., & Bhutada, S. (2023). Student engagement prediction in online session. *International Journal on Recent and Innovation Trends in Computing and Communication*, *11*(2), 43-47. https://doi.org/10.17762/ijritcc.v11i2.6108

[27] Sharma, P., Joshi, S., Gautam, S., Maharjan, S., Khanal, S. R., Reis, M. C., ... & de Jesus Filipe, V. M. (2022, August). Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. In *International conference on technology and innovation in learning, teaching and education* (pp. 52-68). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-22918-3_5

[28] Tao, X., Shannon-Honson, A., Delaney, P., Li, L., Dann, C., Li, Y., & Xie, H. (2022). Data analytics on online student engagement data for academic performance modeling. *Ieee Access*, *10*, 103176-103186. https://doi.org/10.1109/ACCESS.2022.3208953

[29] Thiruthuvanathan, M. M., & Krishnan, B. (2025). Multitask EfficientNet affective computing for student engagement detection. *Multimedia Tools and Applications*, *84*(18), 19039-19063. https://doi.org/10.1007/s11042-024-19815-3

[30] Xie, N., Li, Z., Lu, H., Pang, W., Song, J., & Lu, B. (2025). Msc-trans: A multi-feature-fusion network with encoding structure for student engagement detecting. *IEEE Transactions on Learning Technologies*, *18*, 243-255. https://doi.org/10.1109/TLT.2025.3530457

[31] Yan, L., Wu, X., & Wang, Y. (2025). Student engagement assessment using multimodal deep learning. *Plos one*, *20*(6), e0325377. https://doi.org/10.1371/journal.pone.0325377