Doctoral Dissertation

Application of Skipped Moving Average in Time-Series Data Processing to Enhance Engagement Estimation in Online Video-based Lectures

ZHENG XIANWEN

Supervisor Shinobu Hasegawa

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

March, 2025

Abstract

Online learning, which gained traction in the 1980s, has expanded access to education by offering flexible, interactive, and effective learning opportunities through web-based platforms and learning management systems. The success of online learning is heavily influenced by factors such as self-motivation, student engagement, and interaction between students and instructors. Increased engagement can boost self-motivation and foster more effective interactions, ultimately enhancing the quality and experience of online education. Therefore, improving learner engagement is crucial in overcoming challenges like the digital divide, limited face-to-face interaction, and issues related to self-motivation.

However, in engagement estimation research, due to factors like the Hawthorne effect, existing public datasets often suffer from class imbalance, with relatively few data points representing low engagement levels. This imbalance presents a significant challenge in accurately training and validating machine learning models for engagement estimation. We introduce an original preprocessing approach called "Skipped Moving Average," which not only preserves the integrity of the original video data but also captures its temporal dynamics and variations to address the imbalance issue.

First, to enrich the existing computer vision features and better interpret learners' facial and body language during online learning, we adopted a series of features that can represent facial and body information. Additionally, to further enhance our input features, we experimented with features such as standard deviation and extreme values. We then introduced our proposed Skipped Moving Average data processing method, which includes selecting an appropriate skipping window based on the current data distribution, as well as how to reasonably choose oversampled data segments using cosine similarity. We also experimented with different normalization methods to evaluate their effectiveness in processing video sequence data.

In the experimental phase, we divided the work into two major parts. Experiment 1 used LSTM and LSTM-FCN models to verify whether the proposed SMA preprocessing method could address the issue of imbalance in the current video sequence data. Ultimately, the combination of Skipped Moving Average Oversampling and Standard Deviation for training and validation produced the best outcomes. For engagement estimation with different labels, it achieved Recall/Precision/F1 scores of 0.462/0.157/0.234 for the low label, 0.449/0.504/0.475 for the high label, and 0.456/0.501/0.477 for the very high label. To further validate our proposed method, we also compared

it with the SMOTE oversampling method, which further demonstrated the superiority of our approach.

In Experiment 2, we used transfer learning to verify that the proposed SMA data processing method could be applied to different datasets. The three datasets used in the experiment had varying sample time spans and were quite irregular. Our proposed method achieved Recall/Precision/F1 scores of 0.635/0.720/0.675 for the low engagement label, which is an improvement of nearly 0.25 in the F1 score compared to the results before applying transfer learning.

In this study, we tackled the challenge of class-imbalanced time-series video data in the context of engagement estimation and detection by introducing a novel approach: Skipped Moving Average oversampling. This approach not only mitigates the effects of class imbalance but also preserves the continuity and authenticity of the time-series data, leading to more precise and consistent results in engagement detection.

Keyword: emotional engagement estimation, time-series data, oversampling, online learning, class imbalances data

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Online learning started to gain traction in the 1980s with the advent of computers and the internet[1]. Both Benson[2] and Conrad[3] point out that online learning has improved access to educational opportunities for learners who are described as both nontraditional and disenfranchised. Additionally, online learning has gained popularity due to its potential for providing more flexible access to content and instruction at any time and from any place [4]. This flexibility is enhanced by its connectivity, allowing for varied interactions and making it a versatile mode of education[5]. In this context, the 1990s saw the introduction of web-based training and learning management systems (LMS), which allowed for more structured and interactive learning experiences online[1, 6]. LMS reinforces the learning process through online classroom environments, enabling students to retain their autonomy, enthusiasm, and motivation[1], and allowing anyone with an internet connection to enroll in courses from top universities for free or at a low cost. Research indicates that the efficacy of online learning is comparable to, if not greater than, traditional face-to-face learning when courses are well-designed[7].

1.1 Challenges in Online Learning

The COVID-19 pandemic in 2020 accelerated the adoption of online learning across the globe[9]. With schools and universities being forced to stop courses, institutions rapidly transitioned to online platforms. This period highlighted the importance of digital literacy, the need for robust online infrastructure, and the potential for online learning to provide flexible, accessible education. While online learning offers many benefits, the rapid shift to online learning during COVID-19 also highlighted several challenges. These include the digital divide, the need for self-motivation and student en-

gagement, and the lack of face-to-face interaction. The digital divide refers to the gap between individuals who have access to modern information and communication technology and those who do not[8]. This divide significantly impacts online learning, particularly in terms of access to devices, internet connectivity, a variance in technological skills among students and instructors, and the availability of support and resources.

Self-motivation is crucial for student engagement, especially in online learning. Students who are self-motivated are more likely to take initiative, stay focused, and persist through challenges, leading to higher levels of engagement[10]. Engaged students often find the learning process more rewarding, which in turn boosts their intrinsic motivation[11]. Activities that capture students' interest and provide meaningful interaction can enhance their motivation to learn[10]. There is a reciprocal relationship between self-motivation and engagement. Motivated students engage more deeply with content, and engaging content helps sustain and increase student motivation[10, 11]. Regarding self-motivation, the reduced instruction in online learning means students may struggle to stay motivated[12]. Additionally, The flexibility of online learning offers students the convenience of not having to consider time and location, but it also has the potential to lead to procrastination[12]. Moreover, due to the impact of COVID-19, students cannot participate in a traditional educational environment[69]. This isolation can lead to feelings of loneliness, which can diminish self-motivation[12]. Student engagement is a crucial factor influencing the effectiveness of online learning[13]. The lack of face-to-face interaction dramatically alters traditional educational dynamics, impacting engagement levels[9]. Additionally, online education's reliance on technology means that technical issues, such as connectivity problems and platform malfunctions, can disrupt the learning experience and reduce engagement[9]. Furthermore, creating interactive and engaging content is essential but challenging, as it requires consistent efforts to capture and maintain students' attention and participation[9, 19]. Addressing these challenges is crucial for enhancing the effectiveness of online learning.

The lack of interaction in online education impacts teaching quality not only by reducing student engagement but also by leading to feelings of isolation[13]. Students may feel less motivated to participate actively in discussions and activities without the immediate feedback and encouragement that in-person interactions provide, potentially resulting in lower academic performance. Providing immediate feedback is another challenge due to less interaction[14]. Online education often lacks the immediacy of feedback that in-person education offers, making quick clarification of doubts and instant responses to queries harder to achieve, which can hinder the learning pro-

cess. Additionally, collaborative learning activities, such as group projects and discussions, can be less effective online due to communication barriers and lack of real-time interaction, which also significantly affects the quality of online education[15].

From the above related research, it is evident that self-motivation, student engagement, and interaction between students and instructors, as well as among students, are interrelated and complementary. Self-motivated students are more likely to actively participate, complete assignments, and engage with course materials, thereby increasing overall engagement. Enhanced engagement can directly impact online educational performance, which in turn influences levels of self-motivation. Interaction between students and instructors can stimulate student enthusiasm, which is essential for maintaining high levels of engagement and motivation. In other words, both self-motivation and interaction in online learning are effective means of maintaining high student engagement. They represent unavoidable challenges in online education but also serve as methods to sustain high levels of participation. Therefore, student engagement is a crucial research topic in the field of online learning. If we can enhance and maintain student engagement, we will have identified a key factor in improving the quality of online learning.

1.2 Challenges in Engagement Researches

Engagement, defined as a state of mind that helps learners feel positive and realize quality learning[9], is crucial for maintaining motivation and connection throughout their educational journey. According to Kage's work[16], engagement directly affects the effectiveness of online learning and self-paced courses. High levels of engagement ensure that students remain actively involved, leading to better academic performance and a more fulfilling learning experience. In addition, engaged students are more likely to absorb course material, perform better academically, and retain information longer. Consequently, fostering student engagement is essential for the success of online education programs.

Research into engagement helps understand how students perform in various educational environments and with different instructional content. It aids in identifying effective strategies for maintaining engagement, such as interactive content, regular feedback, and peer collaboration, which are crucial for online learning environments[11, 12]. By addressing challenges like technological barriers, diverse motivational needs, and the quality of interaction, engagement research provides solutions to improve the overall quality of online education[11, 12].

However, unlike traditional in-person classes, where instructors can gauge student understanding and engagement through body language, facial expressions, and eye contact, online learning lacks this immediate feedback mechanism. Instructors find it challenging to assess students' states and levels of engagement during online classes, making it harder to adjust teaching strategies in real-time based on student responses[17]. This limitation can impact the effectiveness of teaching and the overall learning experience in online education. Moreover, the difficulty in gauging learner engagement in online learning prevents instructors from adjusting their teaching strategies to match students' current understanding[17, 18]. This misalignment can make the content seem either too boring or too difficult for students. As a result, learners can quickly lose their engagement, making it challenging to maintain a high level of participation and interest in online classes. Furthermore, empirical studies[11, 19, 20, 21] highlight that learners often have difficulty maintaining a consistent level of engagement, in part due to limited interaction opportunities and a lack of diverse, compelling engagement strategies. Besides, students have different motivational drivers, so personalization in online learning is a crucial development direction. Due to individual differences, understanding how to get student engagement in various online learning scenarios has become increasingly important.

The aforementioned issues in engagement research are all tied to understanding how to estimate and enhance high levels of learner engagement in online learning. Research on engagement is crucial for improving learning outcomes in online education. It helps in identifying effective strategies for adjusting instructional content, facilitating collaboration and feedback between students and instructors, and providing solutions to enhance the overall quality of online education. By understanding how to maintain high levels of engagement, instructors can better tailor their teaching methods to suit students' learning levels and keep them motivated and interested throughout the learning process. Therefore, it is essential that instructors understand and obtain learner engagement in online learning. Although many studies using external devices have achieved some success in estimating student engagement[30, 34, 35, 36], challenges remain in online education environments due to the nature of distance/online learning. Specifically, accurately capturing student engagement without disrupting learners, while staying within budget constraints, continues to be a significant challenge in the field of engagement research.

1.3 Challenges in Engagement Estimation

Estimating and detecting low learner engagement during online learning is a critical challenge for providing appropriate support to students. This difficulty arises from the lack of physical cues and real-time feedback typically available in online classroom settings. To address these issues, researchers have proposed various machine learning approaches to estimate learner engagement, treating it as a complex and challenging task[22]. Such research typically utilizes public datasets for engagement estimation and detection, such as "in the wild"[23] datasets or DAiSEE[24]. The use of these datasets is due to the high cost of constructing datasets annotated with engagement during training and the difficulty of making fair performance comparisons with closed datasets. These datasets, collected in natural settings, present challenges such as estimating engagement from low-illumination face images and dealing with facial occlusions, which complicate accurate engagement detection.

Additionally, as known from the Hawthorne effect, participants may alter their behavior and maintain good engagement when they are aware they are being recorded as part of an experiment[25]. Consequently, such datasets often suffer from class imbalance, with relatively small amounts of data reflecting low engagement levels. This imbalance poses a significant challenge in accurately training and validating machine learning models for engagement estimation. Overcoming these issues is crucial for developing reliable methods to detect and address low learner engagement in online learning environments. Because of the high complexity of low-engagement data[26, 27], it is challenging to model these minority classes during the machine/deep learning process. As a result, minority data are often not effectively classified due to the influence of majority data[26], which affects the accuracy of engagement estimation in this research area[27].

1.4 Challenges in Class Imbalance for Time-Series Data

In this study, we define emotional engagement as the emotional feedback learners exhibit towards learning content and instructors in an online education setting. Emotional feedback from online learners is a reaction expressed through emotions, facial expressions, or physiological signals in response to external stimuli or activities[60, 63]. It is an external manifestation of an internal psychological state, characterized by dynamic changes and external observability. Engagement shares similar characteristics with time-series

data. In the context of insufficient data, oversampling and data augmentation are key research directions for addressing class imbalance in datasets.

Basic data augmentation methods are typically based on transformation-based techniques, such as those applied in the time domain, frequency domain, and time-frequency domain[81, 82, 83]. These include processes like cropping, flipping, jittering, and Fourier transforms, which achieve effects such as video clipping, frame flipping, periodic motion enhancement, and video signal strengthening. Transformation-based methods for video oversampling and data augmentation also include techniques like frame jittering and noise injection[81, 82, 83]. These methods add random noise or Gaussian noise to the original video, aiming to augment the dataset. However, such methods can result in the alteration of the temporal dynamics of the original video, reducing its authenticity. In cases where excessive noise is added, it may even negatively impact the performance of the model.

Advanced methods for oversampling and augmenting time series data include pattern-based methods and generative methods. Pattern-based methods include techniques such as Dynamic Time Warping (DTW), Pattern Mining, and Time Series Mixing[81, 82, 83]. These methods reconstruct the original video to generate new samples, preserving the motion characteristics of the video while enhancing the diversity of the data. However, these methods also face challenges, such as loss of the original video's authenticity and high computational costs. Generative Methods based on deep learning include approaches like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs)[81, 82, 83]. These methods address issues such as video distortion and the stability of generated samples. Combining GANs with sequence modeling techniques allows high-quality synthetic time series data to be generated, ensuring temporal dependency and distribution consistency. VAEs learn the latent space distribution of the video to produce samples similar to the original, ensuring the stability of the generated videos. However, generative models generally have drawbacks, such as complex training processes, high demands on computational resources and dataset size, and potential limitations in fully capturing the dynamic characteristics of complex time-series data.

Other time series oversampling and data augmentation methods, such as decomposition-based methods, automated data augmentation methods, and trend and seasonality decomposition, often have issues with losing the inherent characteristics of time series data and introducing distortion to some extent[82, 83].

All the aforementioned data oversampling and augmentation methods have issues, such as the loss of time-series characteristics, the introduction of noise that affects the training models, and the reduction of the original

data's authenticity. Model-based generative methods also have high training complexity and significant computational resource demands. Research in engagement estimation requires high-quality data that include both the dynamic nature of students' engagement and the authenticity of facial and body features. This is because humans are inherently complex. External facial and bodily expressions are not merely reflections of internal emotions; they are also influenced by personal habits, individual personality, cultural and societal norms, and religious characteristics. If current datasets are augmented by adding excessive noise or sacrificing data authenticity to improve model accuracy, it could limit future research on personalized student engagement and engagement prediction.

1.5 Research Objectives

Based on our investigation of engagement estimation research, the objective of this study is to improve the accuracy of estimating low engagement in learner videos during online learning. To achieve this, we have summarized the following research questions.

- RQ1: How can we address the issue of class imbalance in datasets like DAiSEE?

  Class imbalance often introduces variability in experimental outcomes for engagement classification, making it challenging to accurately detect and estimate different levels of learner engagement in real-time online learning environments. Specifically, low engagement, which is crucial for assessment in real-world contexts, is underrepresented in publicly available datasets. This lack of disengagement data results in prediction inaccuracies, posing a significant challenge to advancing research in this field[71].

- RQ2: How does the proposed method influence the accuracy of engagement estimation?

  Other research challenges include limited data availability and the oversimplification of training samples. These issues often result in suboptimal training outcomes, as models tend to overfit the training and validation datasets. This overfitting negatively impacts the model's ability to generalize effectively, reducing overall performance[43, 72].

- RQ3: Can the proposed method with fine-tuning adapt to different video datasets? In the research on the engagement of learners in online

education, the data is often irregular, as the duration of the classes is not fixed. The length of video data in publicly available engagement estimation/detection datasets also varies[22]. This adds complexity to processing time-series data and directly impacts the accuracy of the model's detection results.

This article introduces an original preprocessing approach called a "Skipped Moving Average", which not only preserves the integrity of the original video data but also captures its temporal dynamics and variation to address this problem. This method aims to mitigate the problems caused by video data imbalances in time-series analysis.

1.6 Structure of the Dissertation

In this dissertation, we first introduce the definition of engagement within the context of our research in Chapter 2, along with a review of related studies in the field of engagement estimation. Chapter 3 provides a detailed explanation of our proposed method for preprocessing imbalanced class video time series data in deep learning, known as the Skipped Moving Average method. Additionally, it describes the body and facial features for deep learning inputs, which are tailored for online learning students. Chapter 4 presents the experiments related to SMA oversampling methods used to address the issue of data imbalance. It includes an introduction to the LSTM and LSTM-FCN models, a comparison of the SMOTE oversampling method in our experiments, as well as the data processing methods and the choices made for various data handling techniques. Additionally, the relevant experimental results are presented. In Chapter 5, we present the experiments conducted on video datasets with different time spans using our proposed SMA method. We also used transfer learning to validate the effectiveness of our approach. This chapter includes details about the experimental models, the parameters used, and the data processing techniques. Finally, we also present the experimental results. In Chapter 6, we discuss in detail the significance of the results from each comparative experiment and outline possible directions for future improvements. Finally, Chapter 7 concludes our research and offers prospects for future work.

Chapter 2

Related Works

2.1 Engagement

2.1.1 Definition of Engagement

Engagement in education refers to the degree of attention, curiosity, interest, optimism, and passion that students show when they are learning or being taught. Defined as a state of mind that helps learners feel positive and realize quality learning, engagement is crucial for maintaining motivation and connection throughout their educational journey[16]. It encompasses their level of participation, willingness to learn, and the investment of their time and energy in the learning process. According to Kage's work[16], engagement directly affects the effectiveness of online learning and self-paced courses, influencing students' motivation, persistence, and overall academic performance.

Engagement is commonly defined in educational terms as encompassing three aspects: cognitive, behavioral, and emotional[17]. Cognitive engagement refers to the depth of thoughtfulness and the willingness of students to expend the effort necessary to understand complex ideas and master difficult skills[7, 28]. This type of engagement involves actively processing information, making connections between new and existing knowledge, and applying critical thinking skills. It is characterized by a student's commitment to learning, problem-solving, and perseverance in the face of challenging tasks. Behavioral engagement is centered around the concept of active participation in both classroom and extracurricular activities[29]. This includes staying focused during lessons, completing assigned work on time, and adhering to an instructor's directions. It reflects a student's involvement in the learning process, demonstrated through consistent attendance, active participation in discussions, and engagement in school-related activities outside the

classroom. Emotional engagement encompasses both positive and negative reactions to instructors, classmates, and learning content[11]. These emotional responses are crucial because they help foster connections with the instructor and classmates, influencing a learner's willingness to participate in learning activities. Positive emotions such as enjoyment and interest can enhance a student's engagement and motivation, while negative emotions like frustration or anxiety can hinder their involvement.

Given the significance of learner-to-instructor interactions and the extended time required to assess cognitive and behavioral engagement, this study focuses on estimating emotional engagement during online learning. Additionally, emotional engagement has the characteristics of immediacy and real-time variability[16] and can be analyzed through students' facial expressions and body information[71]. Therefore, emotional engagement is a key point that is currently more feasible to address among the three types of engagement. We define emotional engagement as the emotional feedback learners exhibit towards learning content and instructors in an online education setting. This includes assessing whether students are actively and positively focused on the learning process.

2.1.2 Approaches in Emotional Engagement Estimation

In recent studies, several prevalent methods have been used to acquire data to analyze student engagement. These methods include analyzing learning log files, tracking clickstream data, administering self-report surveys, using external devices, and employing computer vision techniques.

The use of log files[30], tracking clickstream data[31, 32], and self-report surveys[33] are complementary methods that, when combined, can provide a robust approach to measuring and enhancing student engagement in online learning environments. However, the aforementioned studies are particularly suitable for cognitive and behavioral engagement analysis because of their relatively long periods of relevance. In contrast, analyzing emotional engagement requires methods that can capture real-time and immediate feedback, making it a more dynamic and challenging aspect to study.

The analysis methods for sensor data from external devices, such as EEG, blood pressure, heart rate, or galvanic skin response, have been shown to achieve high accuracy in detecting physiological and emotional states[34, 35, 36]. These methods leverage advanced signal processing and machine learning algorithms to interpret the raw data from sensors, leading to precise measurements of cognitive load, stress levels, and emotional responses. De-

spite their high accuracy, these methods face significant challenges in terms of generalizability. The data collected in controlled laboratory settings often do not translate well to real-world educational environments, where numerous variables and uncontrolled conditions can affect the measurements. Furthermore, the use of such devices in everyday educational settings is impractical due to the need for specialized equipment, the potential discomfort for students, and the complexity of data interpretation. As a result, while promising in research contexts, these sensor-based methods are currently unsuitable for widespread application in real educational settings.

In addition, using keyboard and mouse activity[37] as a measure of online learning engagement has significant limitations. This approach does not apply to learners who participate in online classes using iPads, tablets, or mobile phones. A key feature of online learning is its flexibility with respect to device and location, enabling students to take online courses from virtually anywhere using a variety of devices. By focusing solely on keyboard and mouse activity, these methods overlook a substantial portion of the online learner population and fail to capture engagement data from these widely used devices. This highlights the advantages of computer vision-based methods in engagement research.

Psychological research has shown that facial expressions and body posture are important channels for conveying emotions and thoughts[38, 39]. Computer vision can analyze visual inputs, such as facial expressions and gaze direction, to infer a learner's engagement level. These methods are device-agnostic and can be applied across various platforms, including tablets and smartphones. Moreover, computer-vision-based approaches can provide richer and more nuanced insights into learner engagement by capturing subtle cues that traditional activity tracking might miss, such as micro-expressions and changes in attention. This makes them particularly suitable for the flexible and diverse environments characteristic of modern online learning. Consequently, studying external expressions constitutes an important approach to estimating/detecting learners'emotional engagement. In this study, we adopt a computer-vision-based approach to extract external features that enable the analysis of learners'engagement.

2.2 Computer-Vision-Based Features

In computer-vision-based studies of engagement in online learning, several features are commonly analyzed to assess student engagement. The primary features include facial expressions, gestures, posture, and eye movements. For facial-expression recognition and engagement estimation, techniques such as

Action Units (AUs)[40], Local Binary Patterns (LBPs)[41], and Histogram of Oriented Gradients (HOGs)[42] are widely used and have achieved notable success in related research. Action Units (AUs) are a concept from the Facial Action Coding System (FACS), which breaks down facial expressions into individual components related to muscle movements. These units help in identifying specific emotions and engagement levels based on the combination of activated AUs. Local Binary Patterns (LBPs) are used for texture description and are effective in recognizing facial expressions by analyzing the texture and appearance of different regions of the face. Histogram of Oriented Gradients (HOGs) captures the gradient orientation of facial features, which is useful for detecting subtle changes in facial expressions.

Despite the success of these methods, relying solely on facial expression features has limitations. Facial expressions can sometimes be ambiguous or not fully indicative of a student's engagement level. For instance, a student might show a neutral or bored expression while still being mentally engaged with the material. To address these limitations, some studies have incorporated gestural and postural features into the engagement detection process[17, 43]. Gestures, such as hand movements and body language, can provide additional context about a student's engagement. For example, frequent nodding can indicate understanding and agreement, while fidgeting might suggest restlessness or distraction.

Chang et al.[43] conducted a study utilizing OpenPose to track detailed information about head, body, and hand movements, capturing dynamic changes in subjects'body postures. OpenPose[44], a real-time multi-person detection library, provided precise tracking of body parts, allowing the researchers to measure engagement through physical activity. Specifically, the frequency of hand appearances and the distance between the nose and neck were used as indicators of hand and body movements, respectively. Their findings indicated that fewer restless movements correlated with higher engagement intensity, while more restless movements suggested lower engagement intensity. Regarding eye information, some studies have used eye trackers and existing libraries such as OpenFace to gather data on eye movements[17]. Eye trackers, though highly accurate, were excluded from our study due to their nature as external devices, which could introduce variability and inconvenience in real-world settings. Instead, OpenFace was used to obtain eye information[17], focusing on fixed parameters provided by the library[73]. Additionally, it has been shown in various studies[45] that specific eye-related actions, such as brow raising, brow lowering, and eyelid tightening, are strongly correlated with learner engagement. These micro-expressions offer valuable insights into a learner's cognitive and emotional state, which can be critical for accurately assessing engagement levels.

Based on our preliminary research[48, 49] and related psychological studies[38, 39], distinct patterns in body posture have been revealed, correlating with varying levels of learner engagement. When engagement is low, the learner's body appears tightly closed, with limbs and torso drawn in. As engagement increases, the body posture becomes more open and stretched, indicating a higher level of alertness and involvement. Similarly, head poses vary with engagement levels; a tilted head is often seen at low engagement, while an upright head with a serious expression is common at high engagement. These observations suggest that relying solely on the frequency of hand and body movements does not provide a complete picture of student engagement. In online learning environments, students exhibit a variety of actions that can signal their level of engagement. These include leaning toward or away from the screen, turning their bodies, shoulders, and faces, and using their hands to support their face or adjust their hair. Such diverse behaviors necessitate more sophisticated computer-vision-based features to accurately reflect students' engagement. Advanced techniques in computer vision can address these complexities. For instance, combining the analysis of posture, gesture, and facial expressions provides richer features for assessing engagement. However, existing facial expression recognition libraries, such as OpenFace, provide fixed features that lack flexibility. An important research direction in engagement estimation is analyzing the relationship between facial and bodily features and learners' engagement. Relying solely on existing feature extraction libraries may hinder exploring the relationship between external expressions and internal psychological activities. Moreover, the features of emotional engagement may differ from those used in traditional facial expression recognition. Therefore, building on existing feature extraction research, we adopted facial and bodily expression features that are hypothesized to be strongly correlated with engagement estimation. Utilizing tools like OpenPose to track head, body, and hand movements, and capturing nuanced facial key points offers a more holistic view of engagement. We can select features that better align with our research, such as brow raising, brow lowering, and eyelid tightening, which are strongly correlated with emotional engagement states.

This comprehensive approach ensures that various physical and facial indicators are considered, leading to more accurate and effective measurements of student engagement.

2.3 Dataset

Due to the specificity and complexity of measuring learner engagement, datasets in this area of research require high-quality labels, extensive video durations, large data volumes, and detailed descriptions of learning content. These stringent requirements pose significant challenges. Many successful studies have relied on non-public datasets, which are often curated under controlled conditions and tailored to specific research objectives. However, collecting such large-scale datasets that meet all these criteria within a short time frame is not only challenging but also costly. Public datasets are essential for advancing research in learner engagement because they provide a shared resource that researchers can use to validate findings, compare methods, and build upon each other's work. Despite their importance, maximizing the utility of these public datasets poses significant challenges.

Specific literature reviews[17] identify several publicly available and annotated datasets for engagement estimation and detection using computer vision. The most commonly utilized public datasets include "in the wild" datasets and DAiSEE (Dataset for Affective States in E-Environments). These datasets have been extensively used in various related studies, serving as foundational resources for developing and testing engagement detection models.

Kaur et al.[23] developed a new "in the wild" dataset, published in 2018, featuring video recordings of participants watching stimulus videos. This dataset includes 264 videos, each about five minutes long, with engagement levels labeled by a team of five annotators. The videos were captured using a Microsoft Lifecam wide-angle F2.0 camera at the other end of a Skype video call, simulating real-world conditions like frame drops, network latency, and interference. The dataset comprises 91 subjects (27 females and 64 males) recorded in various settings such as computer labs, dorm rooms, and open spaces. Engagement levels in the videos are categorized from 0 to 3: (0) not engaged, (1) less engaged, (2) engaged, and (3) highly engaged.

Gupta et al.[24] introduced the DAiSEE dataset, which comprises video recordings of learners engaged in online courses, annotated with crowdsourced engagement labels. To collect this data, a high-definition webcam mounted on a computer was used to capture the students' states as they viewed online learning content. The dataset includes 112 participants of Asian ethnicity, with 32 females and 80 males, aged between 18 and 30. It contains 9,068 video snippets, each 10 seconds long, recorded in six different locations under three varying lighting conditions to mimic the diverse environments students might experience during online learning. Each video snippet is assigned a unique identification number and labeled with engagement, frustration, con-

Figure 2.1: Examples of video data from the "in the wild" dataset [23]. The level of engagement increases from left to right.

fusion, and boredom levels. However, in this research, only the engagement labels were utilized, categorized into four levels: (1) very low, (2) low, (3) high, and (4) very high. This dataset aims to reflect the real-world variability in online learning settings, providing a robust foundation for studying learner engagement.

Table 2.1: Overview of "in the wild" [23] and DAiSEE [24] datasets.

Dataset	Subjects	Video Snippets	Snippets Time	Total Time
"in the wild"	78 (male/female 53/25)	197	5 min	59,100 s
DAiSEE	112 (male/female 80/32)	9068	10 s	90,680 s

Table 2.1 compares the basic information of the "in the wild"[23] and DAiSEE[24] datasets. The DAiSEE dataset has several advantages, including a larger number of subjects, greater data volume, and a longer total duration. A key distinction between the two datasets is that the "in the wild"dataset consists of five-minute videos, whereas the DAiSEE dataset comprises 10-second videos. For estimating engagement in online learning, which can fluctuate from moment to moment, it is crucial to derive meaningful insights from short video snippets. This makes the DAiSEE dataset particularly valuable for capturing the dynamic nature of learner engagement.

The "in the wild" [23] and DAiSEE [24] datasets are commonly used public datasets for engagement estimation and detection research. However, these datasets have some limitations. Both datasets are based on participants from a single race, and there is a gender imbalance between female and male participants. Additionally, the DAiSEE and "in the wild" datasets are labeled through crowdsourcing. The DAiSEE dataset was annotated using votes from 10 different annotators on CrowdFlower for each video clip, classifying them into four affective labels: boredom, confusion, engagement, and

Figure 2.2: Examples of video data from the DAISEE dataset [24]. The level of engagement increases from left to right.

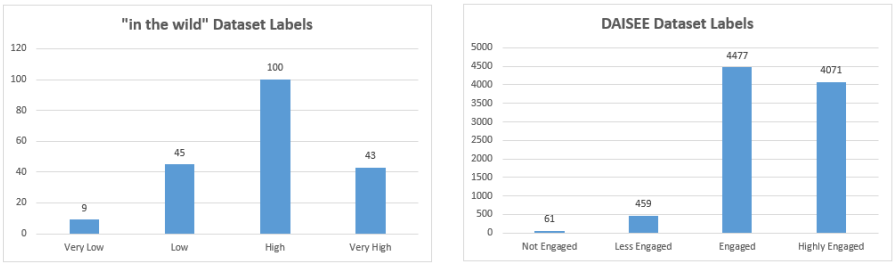

Figure 2.3: Distribution of engagement levels in the “in the wild”and DAiSEE datasets. (We have used the number of downloaded data [75, 76] as the basis for our study).

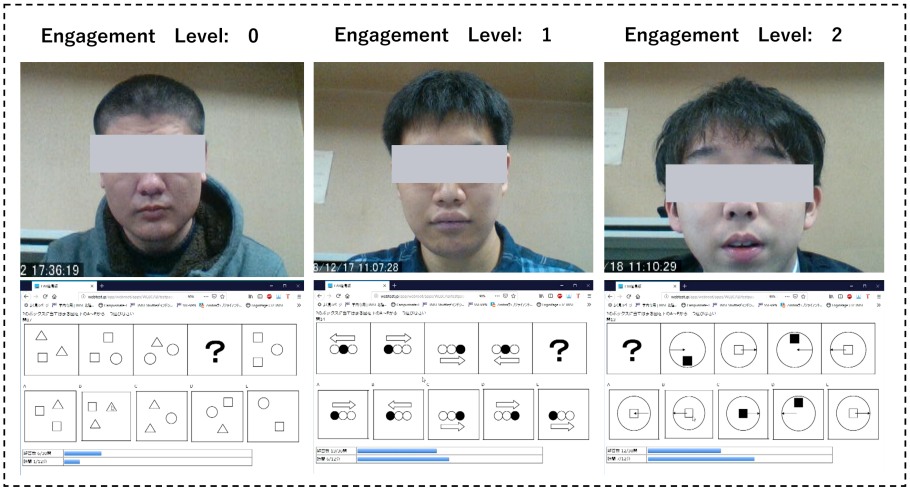

Figure 2.4: Examples of video data from the newly created dataset [72]. The level of engagement increases from left to right.

frustration, with applicability scores ranging from 0 to 3 [84]. The dataset was then split into training, validation and test sets. Each video collected "in the wild" was annotated by five annotators and classified into four levels, ranging from 0 (complete disengagement) to 3 (high engagement). The dataset was divided into three parts, similar to the DAISEE dataset [84]. The crowd-sourcing data annotation method may result in ambiguous labeling due to variations in annotators' cultural backgrounds, knowledge, and subjective judgments, which can introduce bias. Additionally, annotators often lack the specific expertise required for complex tasks like engagement estimation, further contributing to ambiguity in the labeled data[17]. Therefore, crowd-sourcing leads to another limitation: the ambiguity in labeling frames with the appropriate engagement levels[17]. Ambiguity in labeling often arises from the lack of clear guidelines on how to map facial indicators to various affective states or engagement levels of online learners. Another issue with the current datasets is data bias. Figure 2.3 illustrates the distribution of the four engagement levels across the 'in the wild' and DAiSEE datasets. Both datasets exhibit significant class imbalance, particularly in the number of data points for each engagement level. Specifically, instances of low engagement are notably underrepresented.

Given the limitations of existing engagement research datasets, we created a time-series dataset[48, 72] of online tasks involving 19 participants. As shown in Figure 2.4, the videos were recorded using built-in PC cameras during the test-taking process. The online learning content comprised the Cognitive Assessment Battery (CAB) test, which assesses cognitive speed/attention,

episodic memory, visuospatial functions, language, and executive functions. Participants completed up to 30 questions within a 12-minute timeframe. Due to varying individual speeds, the recorded video lengths differed among participants. Upon completing the CAB test, participants submitted self-reports regarding their mental state to confirm engagement levels. The self-reports from participants serve as subjective feedback, allowing them to directly reflect their true psychological states, aligning with their actual emotions and cognitive conditions. This also avoids potential biases that could arise from relying solely on external observations, providing a fundamental guarantee for the engagement data labels. At the same time, self reporting may lead to unstable results due to factors such as social desirability bias, recall bias, and interference from learners' subjective consciousness. To ensure the accuracy of engagement labels, we combined self-reports from the 19 participants with external observations from several study members. As a supplement to external observations, these self-reports also ensure that certain internal details and feelings might not be captured through external observation. This dataset captures the regularity of engagement changes and addresses the need to evaluate the effectiveness of our proposed methods with videos of different lengths.

2.4 Architectures in Emotional Engagement Estimation

In this section, we review previous studies that utilized class-imbalanced datasets to estimate or detect engagement through computer-vision-based methods. Chang et al.[46], in their 2018 study on the "in the wild"dataset, proposed an ensemble framework that combines three cluster-based conventional models and an attention-based neural network (NN) model enhanced with heuristic rules to predict learners' engagement levels while watching Massive Open Online Course (MOOC) videos. Their study applied regression techniques to represent the classification task, reporting class-wise mean square error (MSE) results for engagement levels 0 to 3, which were 0.263, 0.079, 0.032, and 0.136, respectively. The model performed best at Level 2 with an MSE of 0.032 and worst at Level 0 with an MSE of 0.263, primarily due to the imbalanced nature of the dataset that favored the majority classes.

In their 2022 study using the DAiSEE dataset, Villaroya et al.[47] developed an automated engagement detection system that leveraged facial features such as head position, gaze direction, facial expressions, and the distance from the user to the recording RGB camera. The system was primarily

built using the Random Forest algorithm. The classifier's evaluation yielded F1 scores of 0.671, 0.742, 0.890, and 0.860 for engagement levels ranging from very low to very high, respectively. They estimated engagement from shorter decomposed video segments rather than 10-second videos. The discrepancy in results may be due to the unbalanced nature of the DAiSEE dataset used in their study.

Dresvyanskiy et al.[27] utilized a variety of augmentation and class-balancing strategies along with a fusion of emotion-based and attention-based deep embeddings to create a reliable engagement recognition system using facial imagery. They modeled these fused features over time and introduced a novel baseline metric, advocating for performance assessment using the unweighted average recall (UAR) metric. The model's overall performance achieved an accuracy of 39.02% and a UAR metric of 44.27%.

In the aforementioned investigations and review papers[17], class imbalances and insufficient data emerged as common challenges across many studies. Although Villarroya's results are better in low engagement estimation, their study's input is static images. Engagement is characterized by being dynamic and continuously changing[16], so capturing students' facial and bodily expression variations during online classes is also fundamental for future predictive research. Time series features can capture information about an object's actions, changing trends, and temporal patterns in videos[78]. In tasks requiring a deep understanding of time-varying patterns, time series features provide more meaningful information compared to static features. However, utilizing temporal models (such as RNNs, LSTMs, and Transformers) is necessary to capture time dependencies. These models require more complex training processes and demand larger datasets[79]. Additionally, if the extraction of temporal features in videos is inaccurate or affected by noise (e.g., jitter, motion blur), it may impair the model's ability to learn critical temporal information[80]. Static images are temporally independent and are less affected by noise[79]. Additionally, current research on feature extraction and classification models for static images is generally more advanced. While classification based on static images often achieves higher accuracy than time series data, it cannot reflect critical information such as motion trajectories and behavioral dynamics in videos, which are essential for engagement estimation research. Considering the dynamic nature of engagement, relying solely on static images is insufficient for future research. Therefore, while research on engagement classification using images from videos has shown good results, studies on time-series classification have even greater long-term importance.

To address class imbalances issue in our preliminary experiments[48, 49], we employed long short-term memory (LSTM) and quasi-recurrent neural

network (QRNN) sequence models to estimate engagement using time-series facial and body key point information. Utilizing the DAiSEE dataset, we combined very low and low engagement into a single label to mitigate class imbalance. The LSTM model achieved accuracies of 0.050, 0.740, and 0.410 for the three engagement levels, while the QRNN model achieved accuracies of 0.000, 0.930, and 0.070. The F1 scores for the three engagement levels were 0.090, 0.640, and 0.470 for the LSTM, and 0.000, 0.660, and 0.120 for the QRNN, respectively. These results highlight the challenges and potential strategies for improving engagement estimation in class-imbalanced datasets.

We replicated the study by Ai et al.[50], converting their regression results into classification outcomes. Ai et al. proposed an advanced end-to-end framework, Class Attention in Video Transformer, to predict engagement intensity. This architecture relies on self-attention between patches and class attention between class tokens and patches. To address the issue of insufficient training samples, they introduced a binary order representative sampling method, which significantly improved the model's predictive capabilities. The study achieved state-of-the-art mean squared error (MSE) scores of 0.049 for the "in the wild" dataset and 0.037 for the DAiSEE dataset. In our replication, after converting regression results into classification and merging very low and low engagement labels, we obtained classification accuracies of 0.571, 0.789, and 0.667 for the "in the wild" dataset, with F1 scores of 0.696, 0.682, and 0.690, respectively. For the DAiSEE dataset, the classification accuracies were 0.068, 0.732, and 0.421, with corresponding F1 scores of 0.122, 0.625, and 0.489. These results underscore the effectiveness of the proposed framework in predicting engagement intensity across different datasets.

Considering the characteristics of the dataset, the results for the "in the wild"dataset were relatively good even for low engagement levels, whereas the outcomes for the DAiSEE dataset were not as favorable. This aligns with the findings from related studies[17, 43, 47, 50] and corroborates the observations from our preliminary experiments[48, 49], as shown in Table 2.2. These findings underscore the persistent challenge of class-imbalanced data in recent research. This ongoing issue highlights the need for further investigation and the development of robust solutions in this area.

Table 2.2: Preliminary experiments and reproduced results from [33].

Engagement Label	**Dataset**	**Low (Recall/F1)**	**High (Recall/F1)**	**Very High (Recall/F1)**
LSTM [31, 32]	DAiSEE	0.050/0.090	0.740/0.640	0.410/0.470
QRNN [31, 32]	DAiSEE	0.000/0.000	0.930/0.660	0.070/0.120
LSTM [33]	DAiSEE	0.068/0.122	0.732/0.625	0.421/0.489
LSTM [33]	"in the wild"	0.571/0.696	0.789/0.682	0.667/0.690

Chapter 3

Proposed Methods

3.1 Sampling Method

The issue of class imbalance is a significant challenge in many research areas, particularly in machine learning and data analysis. Resampling techniques are commonly employed to balance datasets and mitigate this problem. Oversampling techniques, in particular, are used to increase the representation of the minority class by replicating existing instances or generating new ones, thereby achieving a balanced dataset[51]. One widely used method is the Synthetic Minority Over-sampling Technique (SMOTE)[52]. SMOTE works by generating synthetic samples for the minority class based on the feature space similarities between existing minority instances. While effective, SMOTE has several disadvantages, such as the potential for oversampling uninformative samples, introducing noise, and the indiscriminate selection of neighbors, which can negatively impact the model's performance[53].

To address these issues, we propose a novel oversampling method tailored for video time-series data, called "Skipped Moving Average"[70]. A moving average is a statistical technique commonly used to smooth time-series data by calculating the average of subsets of data points within a fixed-length sliding window[85]. It effectively reduces noise and highlights underlying trends in the data, which is particularly beneficial for regression analysis and forecasting tasks. However, traditional moving averages may not fully address the specific needs of our research context. To overcome this limitation, we have developed the Skipped Moving Average approach. Inspired by traditional moving averages, this method is specifically modified to enhance dataset balancing by addressing the unique characteristics of video time-series data. The Skipped Moving Average method smooths out fluctuations and reduces noise, leading to more informative and representative

samples for the minority class. This approach is designed to improve the accuracy and robustness of engagement detection models in video-based research. Since we are not reconstructing a video based on SMA, but rather dealing with frame features selected based on SMA, we are not regenerating the video itself. Therefore, by preserving the parameters generated after processing the data a skipped moving average, we not only avoid issues related to personal information leakage but also reduce the computational cost of generating new video data.

3.1.1 Skipped Moving Average and Video Frame Undersampling

The Skipped Moving Average is a data preprocessing technique specifically developed to address the issue of data imbalances in video time-series datasets. This innovative method aims to enhance the quality and utility of video data by reducing redundancy and smoothing fluctuations within the video frames. The process involves applying a moving average to the original video frames, which helps in averaging out the variations and noise that can obscure the underlying patterns in the data. In detail, the moving average is calculated by taking the average of every few frames rather than consecutive ones, hence the term "skipped." This approach helps in maintaining the temporal coherence of the video while effectively reducing the noise and redundancy that often plague video time-series data. The result is a cleaner, more balanced dataset that better represents the true distribution of engagement levels or other metrics of interest. By addressing the class imbalance problem and enhancing the dataset quality, the Skipped Moving Average method can lead to more accurate and robust models, ultimately improving the insights derived from video-based analyses.

Considering the data volume, total duration, and the number of subjects, the DAISEE dataset is more suitable for testing our proposed oversampling method. Therefore, in this section, we use the DAISEE dataset as an example to explain and demonstrate our proposed method.

In the DAiSEE dataset, each video sample spans ten seconds with a resolution of 1920 $\times$ 1080 pixels at a frame rate of 30 frames per second (fps)[24]. This results in each sample containing a total of 300 frames. Given that engagement is understood as a sustained affective state rather than a series of fleeting expressions[16], it may not be necessary to capture data at such a high frame rate. Furthermore, the computational latency involved in processing high-frame-rate data with deep learning models must also be considered[54]. As a result, having 30 fps and 300 frames per ten-second

sample might be excessive both in terms of frequency and time span for real-time estimation of student engagement in online courses.

First, considering that the video duration is 10 seconds, we set the sequence timesteps to 10. To ensure that the sampled data remains an integer multiple, the potential moving average windows that can be used are 2, 3, 5, 6, and 10 frames. These windows correspond to oversampling the low engagement data by 15-fold, 10-fold, 6-fold, 5-fold, and 3-fold sampling rates, respectively. Among these options, averaging 2 frames with a 15-fold sampling rate results in 7,800 samples after sampling, which is too high and leads to excessive oversampling. Conversely, averaging 10 frames with a 3-fold sampling rate results in only 1,560 samples after sampling, which is too low and fails to adequately balance the data. Both extremes would likely reintroduce data imbalance issues. Therefore, we focused on averaging windows of 3, 5, and 6 frames, which correspond to 10-fold, 6-fold, and 5-fold sampling rates, respectively. During our preliminary experiments, detailed in Table 3.1, we found that averaging 3 frames with a 10-fold sampling rate led to slight overfitting in low-engagement outcomes, causing unstable recall and F1 scores during testing. Given the need to achieve a balanced sampling rate, we decided to avoid averaging 6 frames and instead opted for averaging 5 frames with a 6-fold sampling rate. This setting provided a more balanced approach and was used as the moving window value in our study. However, it is important to note that while this study identified the optimal parameters for the LSTM model using the DAiSEE dataset, the 5-frame average window may not be universally applicable. Our approach is designed to identify the best settings under the given conditions, and it can serve as a reference point for other researchers working with similar datasets and objectives. The goal is to provide a methodology that can be adapted and refined based on specific dataset characteristics and research requirements.

Table 3.1: Testing results for different Skipped Moving Average windows with 32-D features.

Engagement	**Low (Recall/Precision/F1)**	**High (Recall/Precision/F1)**	**Very High (Recall/Precision/F1)**
LSTM (3 frames average)	0.295/0.079/0.125	0.490/0.508/0.499	0.381/0.507/0.435
LSTM (5 frames average)	0.346/0.090/0.142	0.523/0.521/0.522	0.373/0.537/0.440
LSTM (6 frames average)	0.192/0.066/0.098	0.544/0.501/0.521	0.354/0.500/0.414

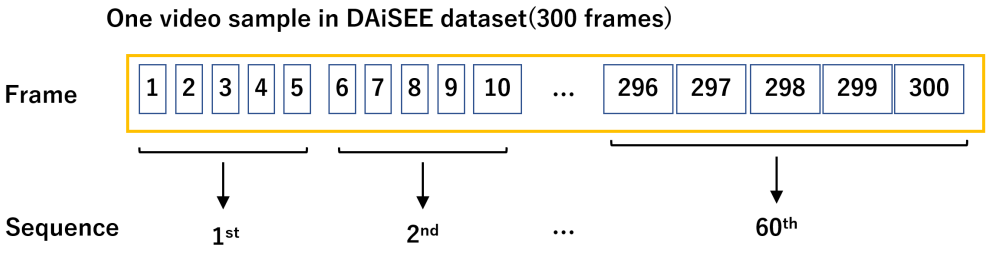

Figure 3.1: Application of the Skipped Moving Average method on a 300-frame video.

We have set the moving average window to 5 frames. By averaging every 5 frames from the 300 frames available in each sample video from the DAiSEE dataset, we obtain 60 average values per sample video. This effectively segments a sample video containing 300 frames into 60 sequences, significantly reducing the amount of data while retaining essential information. The Skipped Moving Average method, as illustrated in Figure 3.1, demonstrates this resampling process applied to video frames from the DAiSEE dataset. By averaging every 5 frames, this method condenses 300 frames into 60 sequences. This approach not only helps mitigate the redundancy present in high-frame-rate videos but also optimizes the dataset for more efficient processing and analysis. This reduction in data size is crucial for real-time engagement estimation, as it decreases the computational load on deep learning models, thereby enhancing their performance and responsiveness.

3.1.2 Average Oversampling Input Videos

In the DAiSEE dataset, there are four levels of engagement labels: (1) very low, (2) low, (3) high, and (4) very high. In Table 3.2, the "Original Labels" row shows the number of video data for the four original engagement labels provided by the dataset. We observed a significant imbalance, with the proportions of very-low and low engagement labels being excessively small. This imbalance posed a challenge for our research, which primarily aims to identify when learners disengage. Given that the four-level classification appeared overly detailed for our purposes, we noted that videos labeled as very-low and low were often very similar[55]. Therefore, we combined the very-low and low labels into a single low-level engagement label, shown as "Relabel" in Table 3.2. This approach simplifies the classification and helps in addressing the data imbalance issue. There are several other examples in

Table 3.2: Engagement labels overview: original labels, data consolidation and oversampling results.

Affective State	**Very Low/Low**	**High**	**Very High**
Original Labels	61/459	4477	4071
Relabel	520	4477	4071
Oversample	2764	4009	3286

the literature of integrating very-low and low labels in this way to improve classification performance[48, 49, 56].

From the previous step, we obtained 60 sequences from each sample video. Given that the sample videos are 10 seconds long, we set the timesteps to 10, resulting in 6 sequences per second. We then sampled 1 sequence from each second to represent the data for that second. After completing this process, each sample video was divided into 6 segments, with each segment consisting of 10 timesteps. To resample the video data, all 6 segments from videos labeled as "low" were retained in their entirety. In contrast, for videos labeled "high" and "very high," one appropriate segment obtained from each second was preserved to form the sample.

Figure 3.2 illustrates the process of oversampling the DAiSEE dataset by segmenting videos into 6 segments. During the data processing phase, some video samples and labels were lost, which resulted in the data presented in the "Oversample" row in Table 3.2. Table 3.2 summarizes the number of original labels, the combined data from the "very low" and "low" labels, and the sample numbers after oversampling. This table provides a clear overview of the distribution of engagement levels in the dataset at different stages: the original distribution, after relabeling, and after applying oversampling techniques to address class imbalances.

3.2 Feature Extraction Method

In related research[17], the terms "emotional" and "affective" are often used interchangeably to describe different aspects of engagement. Affective engagement refers to an emotional response towards learning, which includes feelings such as interest, excitement, and enjoyment in a subject matter[57]. This type of engagement is characterized by the positive emotional connection a learner feels towards the content they are studying, which can significantly influence their motivation and persistence in learning tasks. On the other hand, emotional engagement encompasses a broader range of feelings,

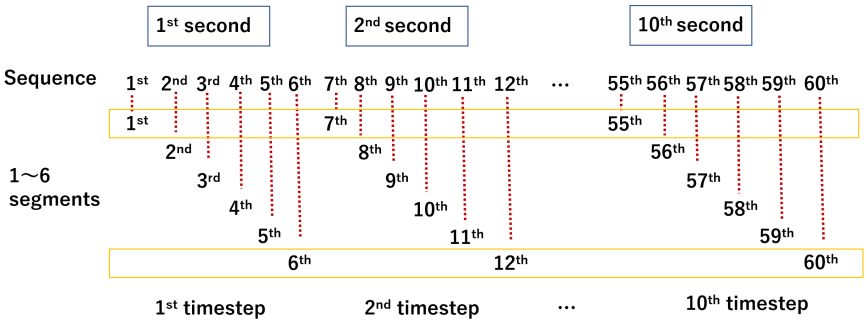

Figure 3.2: Oversampling of the DAiSEE dataset into 6 segments.

both positive and negative, towards instructors, peers, and academic material[11]. This type of engagement includes how students feel about their relationships within the educational environment and their emotional reactions to the instructional content and teaching methods. Positive emotional engagement might manifest as a sense of belonging and enthusiasm, while negative emotional engagement could involve feelings of anxiety, frustration, or alienation. Understanding the nuances between affective and emotional engagement is crucial for instructors and researchers aiming to foster a supportive and motivating learning environment.

3.2.1 Emotional Engagement in Our Study

Bond et al.[58] conducted a comprehensive analysis of 243 studies and identified the five most frequently noted indicators of affective engagement. These indicators, ranked in order of prevalence, are: positive interaction with teachers and peers, enjoyment, a positive attitude towards learning, interest, and motivation. These factors highlight the importance of fostering a supportive and engaging educational environment to enhance students' affective engagement. Conversely, Bond et al. also identified the top five indicators of disengagement, which include frustration, disappointment, worry and anxiety, boredom, and disinterest. These negative emotional responses can significantly hinder a student's ability to engage with the material and achieve academic success. The theory that people's psychological states are expressed through their facial expressions, body language, and the tone and intensity of their voices is widely recognized in the field of psychology[39, 59]. This theory underscores the importance of understanding non-verbal cues in assessing emotional and affective engagement. By recognizing and interpreting these non-verbal indicators, instructors and researchers can gain a deeper under-

standing of students' affective engagement and emotional well-being. This understanding can inform strategies to enhance engagement, reduce disengagement, and create a more positive and effective learning environment.

Moreover, recent research in behavioral science has highlighted the critical role that bodily expressions play in nonverbal communication, more so than was previously recognized[39, 60]. This is particularly pertinent in online learning environments, where many students exhibit limited facial and bodily expressions. Often, students may rest their faces in their hands or cover parts of their faces, which can obscure their facial expressions and make it challenging to analyze their engagement levels based on facial cues alone accurately. In such scenarios, bodily expressions become significantly more important for assessing engagement. Movements such as shifting posture, hand gestures, and overall body orientation can provide vital insights into a student's engagement and emotional state. For instance, a forward-leaning posture may indicate interest and attentiveness, while slumped shoulders could suggest boredom or disengagement. By paying close attention to bodily expressions, instructors can better interpret students' engagement levels, even when facial expressions are not visible or are difficult to discern. This comprehensive approach to analyzing nonverbal communication helps create more effective strategies for maintaining high levels of student engagement and improving the overall learning experience.

Thus, incorporating the analysis of bodily expressions alongside facial expressions provides a more comprehensive understanding of students' emotional and affective states. This dual approach allows for a more accurate assessment of their engagement levels in online learning environments. By integrating both facial and body language cues, instructors and researchers can gain deeper insights into how students are interacting with the material and identify signs of both engagement and disengagement more effectively.

3.2.2 Application of OpenPose in Feature Extraction

In research related to student engagement estimation in online learning, existing libraries such as OpenFace[73] and OpenPose[44] are commonly used to extract computer vision-related information from learners[17]. However, due to the nature of online learning, the information we can capture is limited to the upper body, including the face, shoulders, and arms. As mentioned in Section 2.2, body features play a crucial role in our study, so we require a library that can provide the corresponding body features to support our research. Chang et al.[43] utilized OpenFace in their experiment to capture facial features indicative of engagement, such as head pose and Action Units (AUs). Additionally, to incorporate information regarding changes in a

subject's body posture, they applied two heuristic rules to the outputs of machine learning models. Specifically, they calculated hand movement features by measuring the frequency of hand appearances through wrist features, and defined body fidget features using the distance between the nose and neck, along with the first-order distance delta. However, the heuristic rules applied to adjust the outputs from machine learning models were determined empirically, potentially introducing instability into the results. Additionally, OpenFace provides fixed facial information, which might not offer the level of detail needed to enhance model accuracy in our study. Therefore, the facial keypoints provided by OpenPose give us greater flexibility in feature selection, allowing for improved performance in our models.

Thus, in our study, we utilized OpenPose [61, 62] to extract detailed facial and body key points from learners, enabling us to develop sophisticated computer-vision-based features for analyzing engagement levels during online learning sessions. Using a single tool to capture both facial and body features reduces inconsistencies and enhances the stability of our feature extraction process, thereby minimizing potential errors introduced by combining outputs from multiple separate tools. OpenPose is an advanced real-time multi-person keypoint detection library designed to extract the human body, face, hand, and foot key points from images in real-time, whether from video, webcam, IP camera streams, or locally stored files. Developed by the Carnegie Mellon Perceptual Computing Lab, OpenPose leverages deep learning models to perform human pose estimation by detecting and tracking various keypoints on the human body. Unlike other detection systems that require separate libraries for each type of key point, OpenPose combines the detected body, foot, face, and hand key points into a single output. This unified approach allows for flexibility in choosing any combination of these key points, which can be displayed or saved. Figure3.3 illustrates examples of facial and body key points extracted using the OpenPose library. Figure 3.4 shows examples of key points extracted from the raw footage of participants, as detailed in the reference paper[24]. The images clearly demonstrate that the OpenPose method is effective in obtaining the necessary features from the DAiSEE dataset. This method's capability to capture and integrate multiple key points from various parts of the body ensures that it meets the requirements for detailed and nuanced feature extraction essential for our study.

3.2.3 Facial and Body Features

Psychologists Ekman et al. developed the Facial Action Coding System (FACS) to categorize and understand human emotions such as happiness,

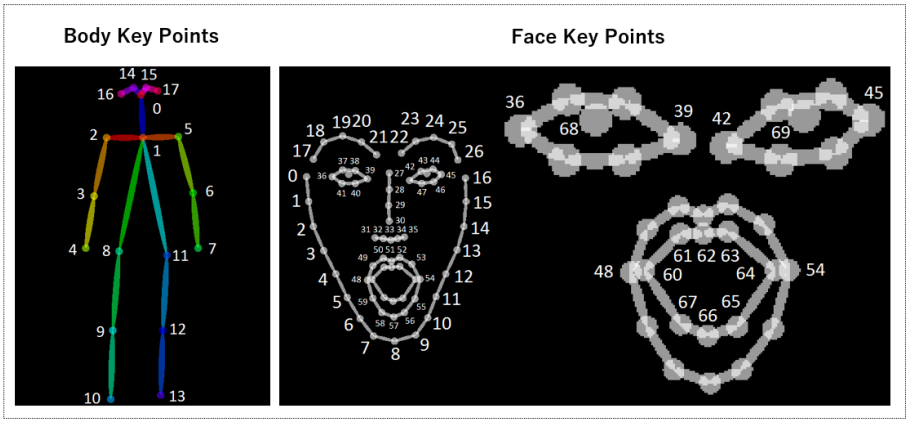

Figure 3.3: OpenOpse key points.

Figure 3.4: Images of key points extracted using OpenPose.

sadness, anger, surprise, fear, and disgust, each associated with specific facial expressions[40]. FACS has become a fundamental tool in psychological research for decoding the complex language of facial expressions. However, it is crucial to note that while there is a significant correlation between facial expressions and emotional/affective states, this relationship can be influenced by various factors, including context, individual differences, and cultural backgrounds[40, 63]. In recent research studies, the association between specific facial expressions and levels of engagement remains imprecise [17]. This ambiguity suggests that while facial expressions are informative, they alone may not fully capture the nuanced levels of engagement in learning environments. Nonverbal communication research underscores the importance of both facial and bodily expressions in conveying emotions, intentions, and attitudes[39]. Facial expressions provide a window into a person's immediate emotional state, but body language offers additional context and depth. In their exploration of body language, Kleinsmith et al. reviewed studies on mapping body features to affective states[39]. They found that the oscillation and movement of body parts, such as the arms, head, shoulders, elbows, and hands, are strongly correlated with the expression of internal emotions. This research highlights the significant role that body dynamics play in conveying psychological states, emphasizing the intricate relationship between physical movements and emotional expressions. These insights underscore the importance of a holistic approach to analyzing engagement and emotions.

Based on relevant research investigations [38, 39, 48, 49, 60], learners' facial and bodily information in engagement estimation studies are typically categorized into the following: gesture and posture, eye information, facial expressions, and additional bodily information. Drawing on the facial and bodily features described in Sections 2.2 and 3.2.1 of related studies, we selected a set of computer-vision-based features to analyze both facial and bodily expressions for estimating engagement levels [48]. These features include eye information, eyebrow and lip shapes, facial rotation angles, head and body posture, the distance between the face and the screen, and body movements.

- Head Posture and Facial Rotation Angles [43, 73]:

Tilt Head

The tilt of the head is calculated using points 27 and 30 to determine the rotation angles of the head in one dimension. This involves analyzing the tilt and orientation of the head, as well as the overall posture of the body.

Turn Head

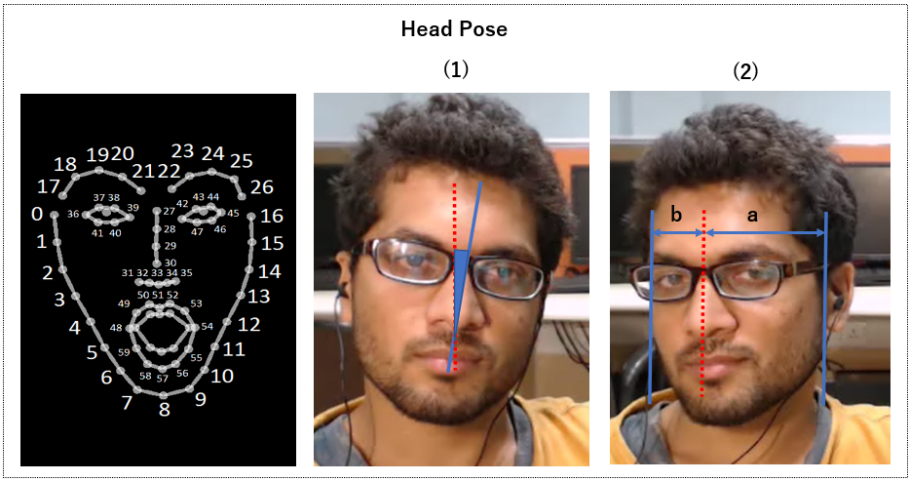

Figure 3.5: Head pose.

The head's turn is determined by subtracting the vector between the right and left face points from the nose point. Specifically, this involves calculating the vector (14-30)-(30-2). If the result is positive, it indicates a turn to the left; if negative, a turn to the right. This measurement in two dimensions provides information about the orientation of the head, indicating where the learner is looking and their level of engagement with the screen.

- Eye Information [17, 43, 45, 73]:

Wink

The wink is calculated by measuring the length between the upper and lower palpebral fissures. Specifically, the distance is measured between points 37 and 41 for the left eye, and points 44 and 46 for the right eye. This provides a two-dimensional analysis.

Eye Movement

Eye movement is determined by measuring the distance between the pupil and the palpebral fissure. For the left eye, the movement is gauged by measuring the distance from point 68 to points 36, 37, 39, and 41. For the right eye, it is measured from point 69 to points 42, 44, 45, and 46. This results in an eight-dimensional analysis.

These metrics include blink rate, gaze direction, and eye movement patterns, which are indicative of attention and focus.

- Lip Shapes [40, 45]:

Open/Close Mouth

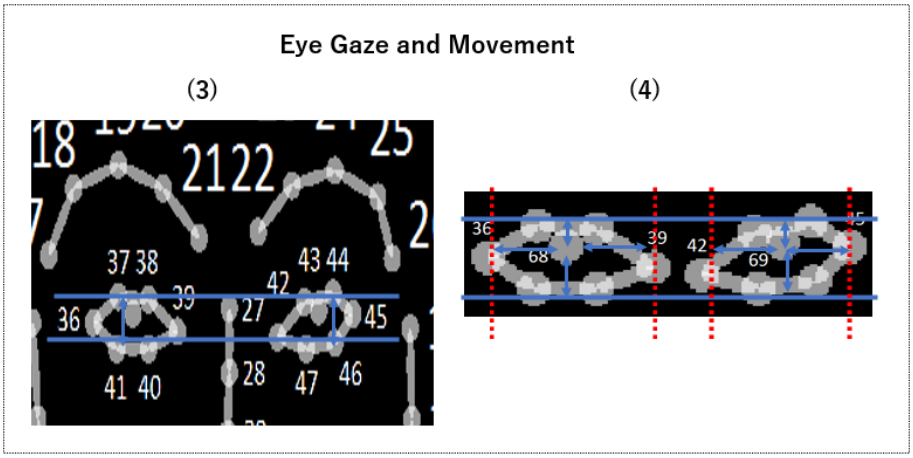

Figure 3.6: Eye Information.

The degree to which the mouth is open or closed is determined by dividing the distance between the upper and lower lips by the width of the mouth. Specifically, the value of (62−66)/(54−48) is used. A larger value indicates an open mouth, while a value of zero indicates a closed mouth. This provides a one-dimensional measurement.

Mouth Angle Radian

The angle of the mouth is measured by the vertical distance from the bottom lip to the corners of the mouth. The calculation (66−(48+54)/2) is used. If the result is positive, it indicates unhappiness, confusion or dissatisfaction; if negative, it indicates a normal or happy expression. This also provides a one-dimensional measurement.

- Body Movement [43, 59, 60]:

Head Movement

The head movement is calculated by determining the standard deviation (STD) of point 30 across one sample video. This is done by subtracting the STD of point 30 from its value in each frame of the video (STD30). This provides a one-dimensional measurement.

Hand and Elbow Movement

The presence or absence of hand and elbow movements in the sample video is noted, with a value of 1 given if the hand or elbow appears and 0 if it does not. This results in a two-dimensional measurement.

The frequency and nature of body movements, such as hand gestures, arm movements, and shifts in sitting position, provide additional context about the learner's engagement and emotional state.

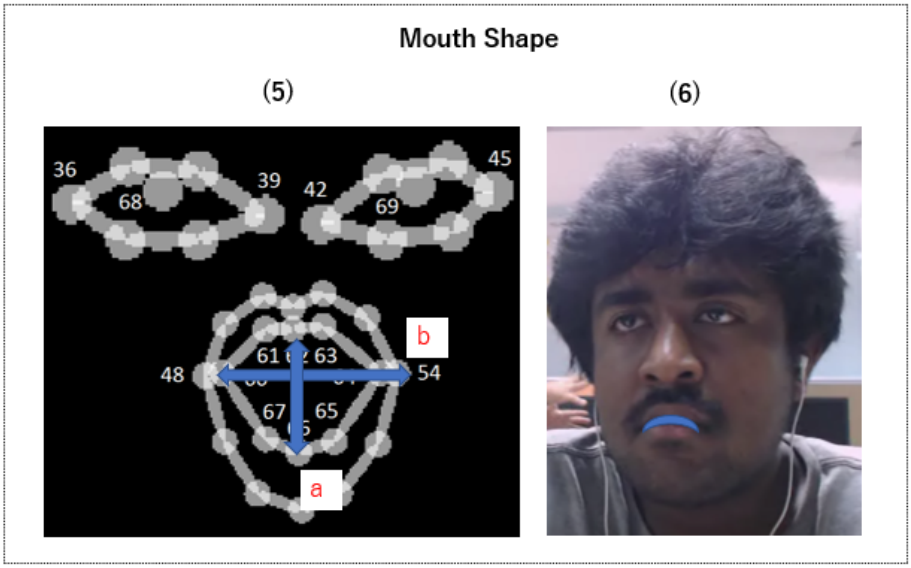

Figure 3.7: Lip shapes.

- Distance Between the Face and the Screen:

Distance from Screen

The distance between the face and the screen is calculated by measuring the length between the two eyes. This provides a one-dimensional metric.

Since the distance between a learner's two eyes is generally consistent across individuals, the variation in inter-eye distance among different learners becomes even smaller when captured in videos. Therefore, we assume that changes in the inter-eye distance reflect variations in the distance between the learner's face and the display screen. The proximity of the learner's face to the screen can be a useful indicator of engagement. Leaning closer to the screen may suggest higher levels of interest and concentration.

In our experiment, we utilized the original body key points of the shoulders, elbows, and hands as an additional body feature. This provided a comprehensive set of 32 dimensions for analysis. The inclusion of these body key points allowed us to measure various aspects of physical engagement, such as the position and movement of the upper body, which can be indicative of a learner's focus and interest. By tracking the shoulders, elbows, and hands, we could analyze gestures and shifts in posture that contribute to a more nuanced understanding of engagement levels.

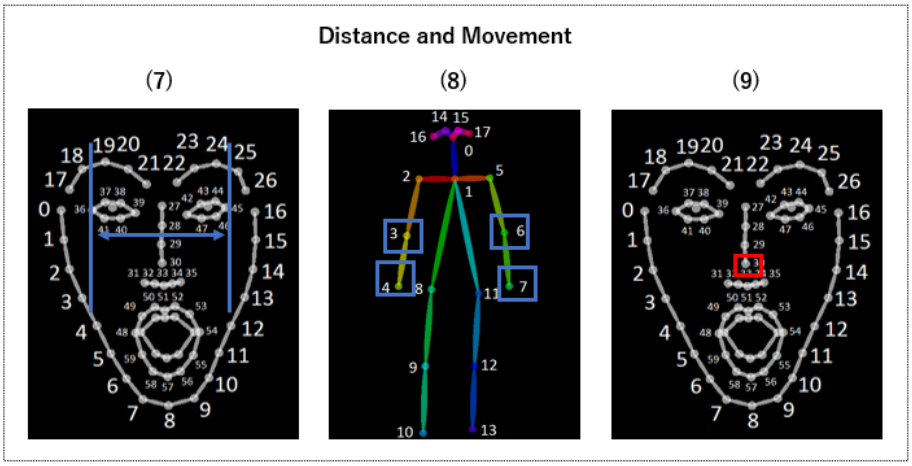

Figure 3.8: Distance and movement.

3.2.4 Additional Features: Standard Deviation and Extreme Values

In the feature section, we adopted relevant features such as eye information, eyebrow and lip shapes, facial rotation angles, head and body posture, the distance between the face and the screen, and body movements. To oversample the existing data, we applied the Skipped Moving Average method to the video samples. This technique smoothed out some data noise but also potentially lost some information about the learners' body movements. However, in online learning, body movements are crucial for assessing engagement. For example, within a specific time frame, larger ranges and frequencies of body movements indicate lower engagement, while smaller ranges and frequencies indicate higher engagement. Simply averaging the input video frames fails to capture the range and frequency of body movements within a unit of time.

Standard deviation (SD) is a crucial statistical measure that quantifies the amount of variation or dispersion in a set of data points. It is the square root of variance and provides a direct insight into the spread of data around the mean. A low SD indicates that data points are close to the mean, showing less variability and more consistency in the data. Conversely, a high SD indicates a wide range of values, suggesting more variability and less consistency. In other words, a lower SD value represents lower engagement levels from the learner, indicating that their behavior or responses are more consistent and show less variation. On the other hand, a higher SD value signifies higher engagement levels from the learner, suggesting that their behavior or responses vary widely. Similarly, the maximum and minimum values within the same window also convey significant information about the learner's body

movements. These critical values can provide insights into the range and extremity of movements, which are important indicators of engagement levels. By incorporating SD, maximum and minimum values into our analysis, we gain a deeper understanding of the temporal dynamics within each video, leading to more accurate and robust models for engagement detection and other analytical tasks. Therefore, we also added the standard deviation, minimum, and maximum as features to enhance the model's performance by accurately reflecting the variability in body movements.

During the skipped moving average data processing stage, for a 10-second video, we used a window of 5 frames to calculate an average value for the video. This resulted in the processed 300-frame video sample being divided into 60 segments. To calculate the standard deviation and the minimum and maximum values, we also used a 5-frame window to ensure consistency. This approach ensures that the periodicity of our proposed Skipped Moving Average data extraction remains the same.

Chapter 4

Experimentation with Skipped Moving Average for Oversampling in Class Imbalanced Datasets

4.1 Purpose

In this stage of the experiment, we aim to verify whether our proposed oversampling method, Skipped Moving Average (SMA), can address the class imbalance in time series video data related to learners' engagement during online learning. Considering the real-time variability of learners' engagement, we will use the DAiSEE dataset to evaluate our method. To further assess the effectiveness of our approach, we will also compare it with the widely used Synthetic Minority Over-sampling Technique (SMOTE)[66].

4.2 Training Method

4.2.1 Application of LSTM in the Training Method

One of the characteristics of engagement is its dynamic and fluctuating nature[16]. Engagement is not a static state; it continuously changes, making it crucial to capture and predict these variations. To evaluate our proposed data preprocessing and oversampling methods, we conducted experiments using a time-series deep learning model, LSTM[64]. Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) architecture designed to effectively learn and remember over long sequences of data. They

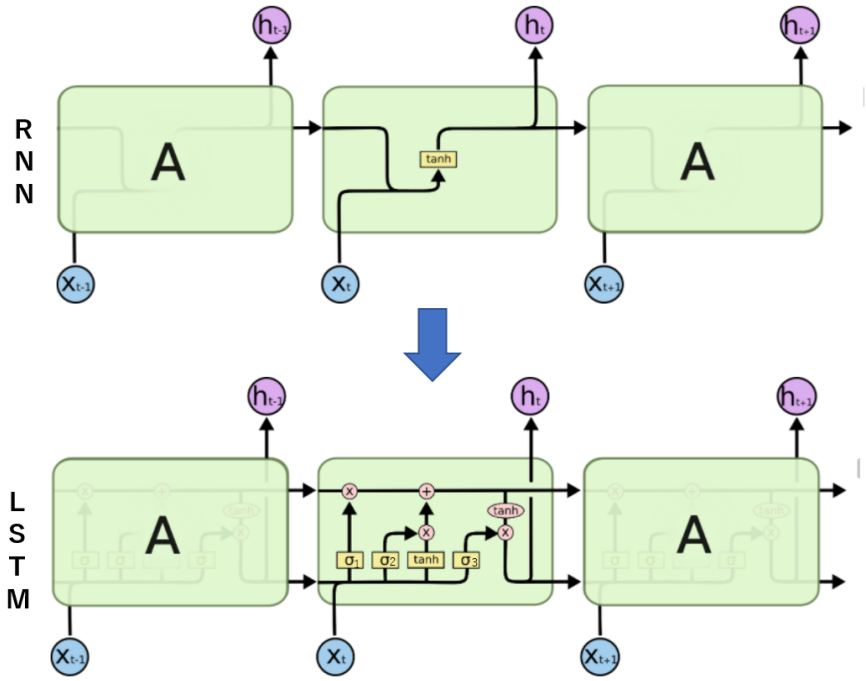

Figure 4.1: Recurrent neural networks and Long Short Term Memory networks[64].

were introduced by Sepp Hochreiter and Jürgen Schmidhuber in 1997 and have since become a popular choice for a wide range of sequence learning tasks due to their ability to mitigate the vanishing and exploding gradient problems commonly encountered with traditional RNNs. This model is well-suited for tasks that involve sequential data and temporal dependencies, making it ideal for analyzing the ongoing changes in engagement levels.

Figure 4.1 illustrates the transition from recurrent neural networks (RNNs) to long short-term memory (LSTM) networks. In LSTM architecture, the key components are the cell state and the gates, which work together to manage the flow of information through the network.

- Cell State: The core concept of LSTM is the cell state, which acts as a conveyor belt that runs through the entire sequence, carrying relevant information throughout the processing of the input data. This helps in preserving long-term dependencies.

- Gates: LSTM networks use three types of gates, each of which plays a crucial role in regulating the information flow:

Forget Gate (σ_1): Decides what information to discard from the cell state. It takes the previous hidden state and the current input and outputs a number between 0 and 1 for each number in the cell state, with 0 representing "completely forget" and 1 representing "completely keep".

Input Gate (σ_2): Decides which new information to add to the cell state. It works in combination with the $tanh_1$ layer, which generates a new vector contributing to the cell state update.

Output Gate (σ_3): Decides what part of the cell state to output. This output is based on the cell state and the input to the output gate.

In the update process, the pink circles represent how the previous cell state is updated by the forget gate's output and the new candidate values generated by the tanh layer.

- Activation Functions:

  LSTMs rely on sigmoid and tanh activation functions to regulate the flow of information within the network. These functions help in controlling the amount of information passed through the gates and into the cell state.

Due to their architecture, LSTM networks are highly effective for various tasks involving sequential data, including classification, processing, and forecasting. Moreover, LSTMs can process other forms of sequential data, including images (when treated as sequences of pixels) and audio or video time-series data, making them versatile tools for handling complex, temporally dependent data. In this study, we leveraged the capabilities of LSTM networks to evaluate our proposed data preprocessing and oversampling methods.

The overall structure of the LSTM model is illustrated in Figure4.2. The proposed LSTM model features a single LSTM layer consisting of 32 hidden units. Following the LSTM layer, a fully connected layer was used to map the LSTM outputs to the desired output space. This architecture leverages the capabilities of LSTM networks to capture temporal dependencies in the data while maintaining computational efficiency.

4.2.2 Application of LSTM-FCN in the Training Method

Karim et al.[65] proposed a hybrid model that combines Fully Convolutional Networks (FCNs) with Long Short-Term Memory (LSTM) networks to enhance time series classification tasks. This innovative approach leverages

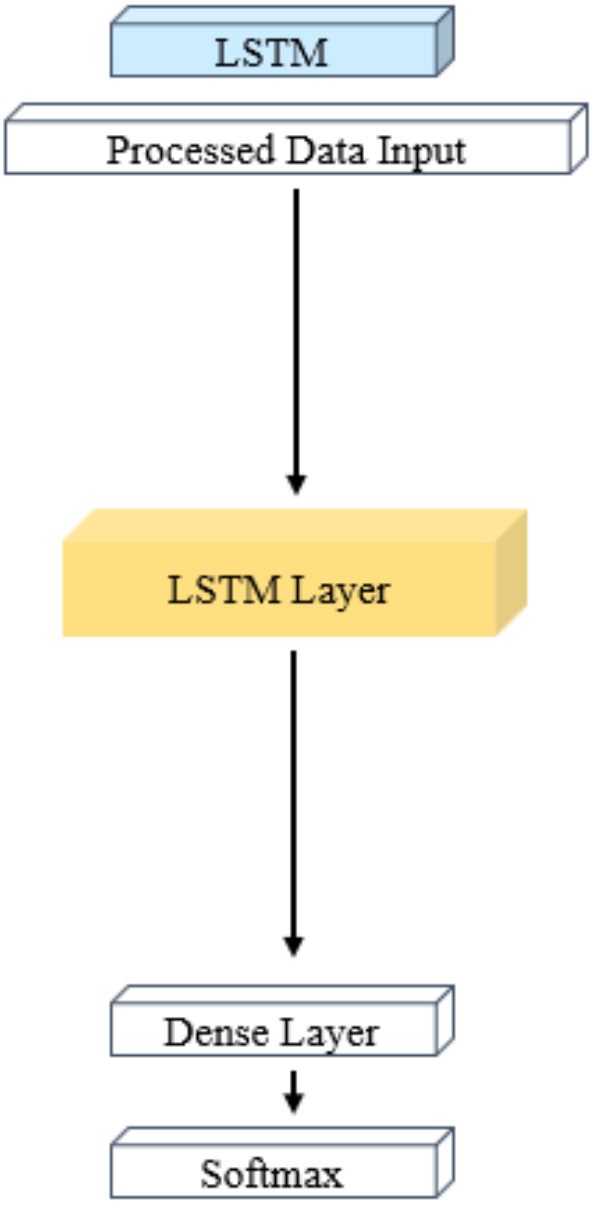

Figure 4.2: LSTM architecture.

the strengths of both FCNs and LSTMs to improve the performance of time series classification.

The proposed LSTM-FCN model combines the strengths of Fully Convolutional Networks (FCNs) and Long Short-Term Memory (LSTM) networks. This hybrid model is composed of two main components: the FCN Component and the LSTM Component.

- The FCN (Fully Convolutional Network) component:

  The FCN part processes the time series as a univariate sequence with multiple time steps, extracting features through convolutional layers. The fully convolutional block consists of three stacked temporal convolutional blocks with filter sizes of 128, 256, and 128, respectively.

  Each block includes:

  1. A temporal convolutional layer.
  2. Batch normalization.
  3. A ReLU activation function.

  Finally, global average pooling is applied after the final convolution block.

- LSTM Component:

  The proposed architecture includes a dimension shuffle layer and an LSTM block. The time series input is first conveyed into the dimension shuffle layer. This layer transforms the time series data, which is then passed into the LSTM block. The LSTM block in the proposed architecture processes the input time series as a multivariate time series with a single time step. The LSTM block consists of either a general LSTM layer or an Attention LSTM layer, followed by a dropout layer. This structure allows the model to capture long-term dependencies and tempor.

- Concatenation Component:

  The output of the global pooling layer from the FCN (Fully Convolutional Network) component and the LSTM component is concatenated and then passed on to a softmax classification layer. This final layer produces the probability distribution across the different classes for the time series classification task, enabling the model to make accurate predictions based on the combined features extracted by both the FCN and LSTM components.

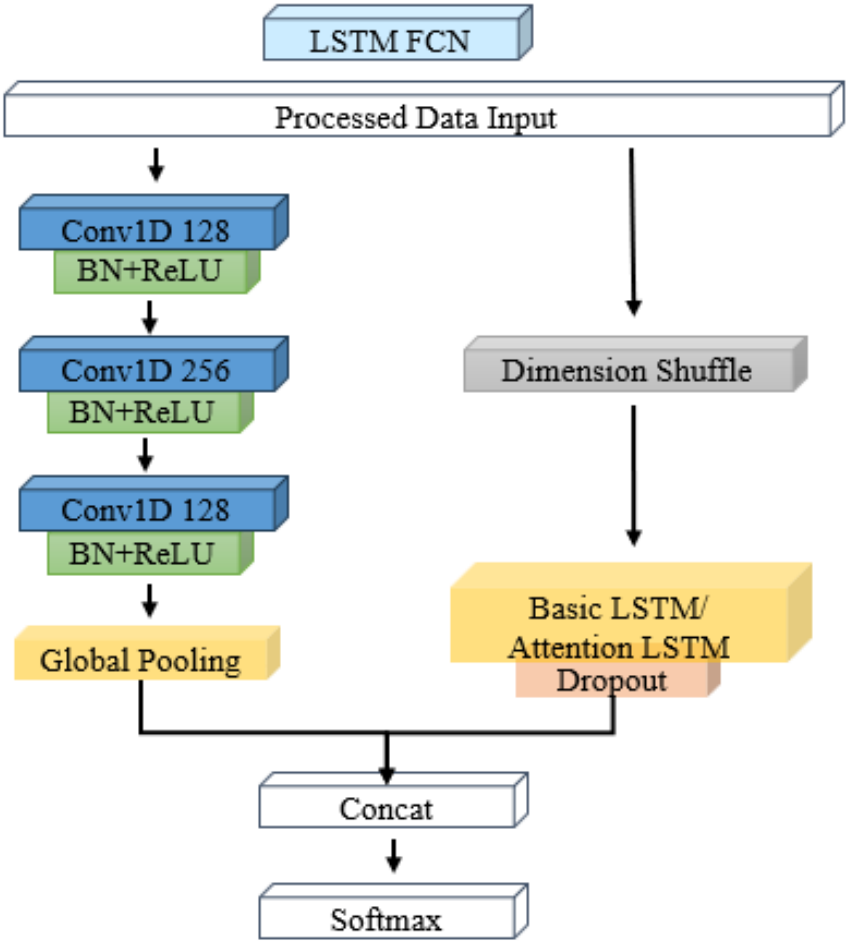

Figure 4.3: LSTM-FCN architecture.

By integrating these two components, the LSTM-FCN model effectively captures both local and long-term patterns in the data, making it highly effective for time series classification tasks. This architecture leverages the convolutional layers' ability to identify short-term features and the LSTM layers' strength in recognizing long-term dependencies, providing a robust framework for analyzing complex time series data.

4.3 Data preprocessing

4.3.1 Skipped Moving Average with Oversampling

In the data processing section, we applied the Skipped Moving Average method for data preprocessing. As described in sections 3.1.1 and 3.1.2, we performed sixfold oversampling on the significantly small low engagement categories in the DAISEE dataset, such as "not engaged" and "very low".

1. Merging Label Data

   Firstly, as mentioned in section 3.1.2, videos labeled as "very low" and "low" in the DAISEE dataset are often very similar [55]. Therefore, we merged the "very low" and "low" labels into a single "low engagement" label. This approach simplifies the overly detailed four-class classification problem and helps address the issue of data imbalance.

2. Undersampling Redundant Frames

   Since the video samples in the DAISEE dataset are 10 seconds long, to preserve the characteristics of the sequential data, we set the step length of each video input sample to 10 time steps. To use the optimal window settings described earlier, we set the moving window for the Skipped Moving Average to 5 frames. After calculating the average values of the video frames, we obtained 60 sequences.

3. Oversampling Video Samples

   From the above averages, it can be seen that the 30 frames per second are reduced to 6 sequences per second. Then, we sample 1 sequence from each second to represent the data for that second. After completing this process, each sample video is divided into 6 segments, with each segment consisting of 10 time steps. This means that an oversampled 10-second video sample is effectively multiplied by 6 times. Additionally, all 6 segments of the videos labeled as "low" retain their sequential information in full, ensuring that the oversampled data remains both authentic and effective.

By combining these steps merging labels, undersampling redundant frames, and oversampling video samples—we significantly improve the quality and balance of the DAISEE dataset. Specifically, the training and validation data distribution after processing is as follows: 2477 samples labeled as “low,”3471 samples labeled as “high,”and 3096 samples labeled as “very high.”From the processed results, it is evident that we significantly increased the amount of data labeled as ”low,” thereby reducing the disparity in the distribution of input data compared to the original dataset.

4.3.2 Random Moving Average with Oversampling

We proposed the SMA method, where a fixed number of video frames per second are selected and averaged as input to the model. Correspondingly, to further validate the performance of the SMA method, we also conducted experiments using random moving average(RMA) to compare diverse settings.

First, we applied the same processing described in Section 4.3.1, 1. Merging Label Data. Second, to ensure consistency in the experimental data settings, we set the step length of each video input sample to 10 time steps. Subsequently, 6 frames were randomly selected per second and averaged as shown in 4.4. Finally, during the oversampling phase, data labeled as ”low” underwent six rounds of random sampling, with the average value calculated for each round. This process increased the data labeled as ”low” to 6 times, ensuring alignment with the previous data settings. The RMA process was applied only once for the ’high’ and ’very high’ label data. The processed data was then used to train and validate the model. The same random moving average process was applied once for the test data.

We obtained the same distribution of training and validation data as with SMA oversampling: 2477 samples labeled as “low,”3471 samples labeled as “high,”and 3096 samples labeled as “very high.”

4.3.3 Application of SMOTE in Our Experiment

To validate the effectiveness of our proposed method, this study compares the Skipped Moving Average (SMA) approach with the widely used Synthetic Minority Over-sampling Technique (SMOTE). By evaluating the performance of classifiers trained with these different data processing techniques, we aim to demonstrate the advantages of the Skipped Moving Average method over traditional oversampling methods like SMOTE.

Chawla et al.[66] introduced the Synthetic Minority Over-sampling Technique (SMOTE), a novel method designed to address the common issue of

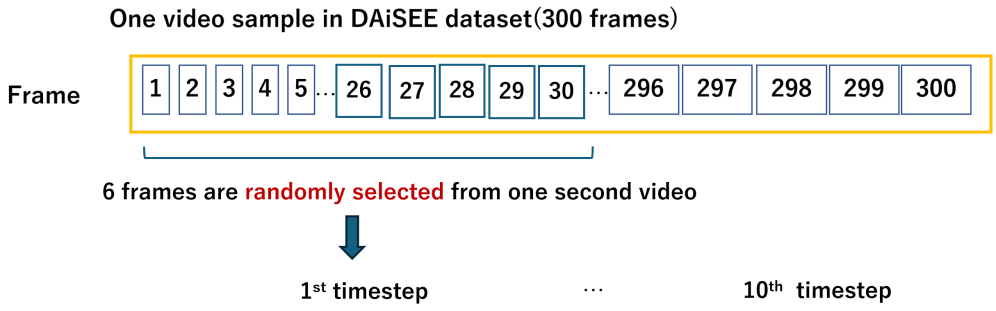

Figure 4.4: Application of the Random Moving Average method on a 300-frame video.

class imbalance in datasets. This imbalance, where the minority class is significantly underrepresented compared to the majority class, is prevalent in many real-world applications and often leads to poor classifier performance. SMOTE aims to improve the performance of classifiers trained on imbalanced datasets, where the minority class is significantly underrepresented compared to the majority class. Instead of simply duplicating minority class examples, SMOTE generates synthetic samples.

They create synthetic examples in a more generalized way by working in the ”feature space” instead of the ”data space.” To oversample the minority class, they take each sample from the minority class and generate synthetic examples along the line segments connecting it to any or all of its k nearest minority class neighbors. Depending on the desired amount of over-sampling, neighbors are randomly selected from the k-nearest neighbors, enhancing the diversity of the synthetic samples.

However, despite its effectiveness, SMOTE has several limitations that need to be addressed. These include the oversampling of uninformative samples, potential noise interference, and the indiscriminate selection of neighbors[53]. These issues can reduce the quality of the synthetic samples and adversely affect the performance of the classifiers.

4.3.4 SMOTE with Oversampling

To further validate our proposed SMA method for processing video sequence data, we compared the results of SMA with the SMOTE oversampling method. The data preprocessing steps and SMA processing method remained consistent, with the only difference being that SMOTE was used for the oversampling portion. The details are as follows:

After completing the 1. Merging Label Data and 2. Undersampling Redundant Frames processing steps, 3. we performed fivefold oversampling on the same sequences that were sampled in SMA. Since SMOTE generates oversampled data points that are very close to the original points in the raw data, we used fivefold oversampling in addition to the raw data to maintain consistency in the data. This approach also ensures the validity of the comparison experiments.

Thus, after applying the same processing steps but with different oversampling methods, we obtained the same distribution of training and validation data as with SMA oversampling: 2477 samples labeled as "low,"3471 samples labeled as "high,"and 3096 samples labeled as "very high."

4.3.5 Selection of Window Period

In the Undersampling Redundant Frames step, we processed the video samples from the DAISEE dataset using a window period of 5 frames, the processed videos had 60 segments per video. Since this is done on a per-second basis, each second of the video was divided into 6 sequences. In the Oversampling Video Samples step, we oversampled the video samples with low engagement labels data distribution. For the video samples with high engagement labels data distribution, we only selected one segment per second to form the input data for the deep learning training model. This approach ensured that the segments with high data distribution maintained the authenticity, temporal sequence, and consistency of the original video data.

However, the undersampled data has 6 sequences per second, meaning there are six segments available for selection. Choosing the appropriate time sequence for sampling is also an issue that requires further investigation. To validate our method, we compared the cosine similarity of different segments for high distribution labels after data preprocessing and oversampling of low label data distribution.

Table 4.1 and Table 4.2 show the cosine similarity results of different segments after data preprocessing and oversampling of low data distribution is shown. These results were calculated for the training data, considering the sequences obtained after applying the Skipped Moving Average for averaging. Among the 6 sets of training segment data, the highest cosine similarity is between the first and second segments, with a value of 0.958. The average cosine similarity of each segment with the other five segments in training data is as follows:

- segment 01 with the other five segments: 0.9262
- segment 02 with the other five segments: 0.9270

Table 4.1: The cosine similarity results for different segments in the training data after data preprocessing and oversampling of low label data distribution.

	01	**02**	**03**	**04**	**05**	**06**
01	1.000	0.958	0.929	0.922	0.918	0.904
02		1.000	0.902	0.933	0.925	0.917
03			1.000	0.930	0.925	0.952
04				1.000	0.944	0.916
05					1.000	0.893
06						1.000

- segment 03 with the other five segments: 0.9276
- segment 04 with the other five segments: 0.9290
- segment 05 with the other five segments: 0.9210
- segment 06 with the other five segments: 0.9164

From the results listed for the average cosine similarity of each segment with the other five segments, it is evident that the fourth segment has the highest similarity. Additionally, as shown in Table 4.1, the highest cosine similarity of 0.958 is found between the first and second segments. Therefore, in the 60 segments obtained from undersampling, we select the first and fourth segments for validation in the non-oversampled data section. The first and fourth segments represent the beginning and middle parts of each second in the video, respectively. This selection is crucial for verifying the processing and evaluation of time-series data.

In the training data, we applied oversampling to the low distribution labels. To evaluate our proposed method and model, we only calculated the input features for the test data and did not perform any oversampling. Therefore, we performed 6 separate samplings on the undersampled sequence data and compared the cosine similarity of the 6 sets of data. In Table4.2, the cosine similarity results of the 6 sets of segment data are listed. The average cosine similarity of each segment with the other five segments in testing data is as follows:

- segment 01 with the other five segments: 0.9866
- segment 02 with the other five segments: 0.9840
- segment 03 with the other five segments: 0.9788

- segment 04 with the other five segments: 0.9824
- segment 05 with the other five segments: 0.9804
- segment 06 with the other five segments: 0.9858

From Table 4.2, it can be seen that the cosine similarity of 0.990 between the first and second segments is the highest, which is consistent with the results of the cosine similarity between different segments in the training data. The segment with the highest average cosine similarity with the other five segments is the first segment in the test data. However, to maintain consistency with the training data, we also selected the first and fourth segments in the test data. This is also to test the effect of selecting the beginning and middle times in a time series on the model results. The results in Table 4.3 are also consistent with the cosine similarity results. The first and fourth segments have the highest ACC.

Table 4.2: The cosine similarity results for different segments in the testing data after data preprocessing.

	01	**02**	**03**	**04**	**05**	**06**
01	1.000	0.990	0.981	0.983	0.991	0.988
02		1.000	0.976	0.984	0.983	0.987
03			1.000	0.981	0.970	0.986
04				1.000	0.977	0.987
05					1.000	0.981
06						1.000

Simultaneously, Tables 4.4 and 4.5 present the cosine similarity results for the data with added standard deviation after data preprocessing and oversampling of low distribution label data. Since only the standard deviation feature was added and the overall data structure did not change, the cosine similarity of the six segments was compared similarly. The average cosine similarity of each segment with the other five segments for average + standard deviation features in the training data is as follows:

- segment 01 with the other five segments: 0.846
- segment 02 with the other five segments: 0.845
- segment 03 with the other five segments: 0.830
- segment 04 with the other five segments: 0.846

Table 4.3: Precision, Recall, and F1 scores for different segments in the testing data after data preprocessing.

Engagement	Acc	Low (Recall/Precision/F1)	High (Recall/Precision/F1)	Very High (Recall/Precision/F1)
segment 01 (Features)	0.445	0.346/0.090/0.142	0.523/0.521/0.522	0.373/0.537/0.440
segment 02 (Features)	0.431	0.372/0.087/0.141	0.498/0.514/0.506	0.366/0.530/0.433
segment 03 (Features)	0.434	0.397/0.106/0.168	0.487/0.518/0.502	0.382/0.492/0.430
segment 04 (Features)	0.458	0.282/0.106/0.154	0.546/0.511/0.528	0.385/0.512/0.439
segment 05 (Features)	0.411	0.346/0.076/0.124	0.502/0.497/0.499	0.322/0.524/0.399
segment 06 (Features)	0.435	0.299/0.090/0.138	0.516/0.506/0.511	0.364/0.495/0.420

Table 4.4: The cosine similarity results for different segments with average + standard deviation features in the training data after data preprocessing and oversampling of low label data distribution.

	01	**02**	**03**	**04**	**05**	**06**
01	1.000	0.927	0.825	0.844	0.854	0.780
02		1.000	0.796	0.838	0.844	0.822
03			1.000	0.841	0.824	0.862
04				1.000	0.873	0.832
05					1.000	0.761
06						1.000

- segment 05 with the other five segments: 0.831
- segment 06 with the other five segments: 0.811

Interestingly, after adding the standard deviation feature, the cosine similarity comparison in the training data also showed that the highest similarity was between the first and second segments, with a value of 0.927. When comparing the average cosine similarity of each segment with the other five segments, the first and fourth segments had the highest average similarities, both being 0.846. This is consistent with our previous selection.

The average cosine similarity of each segment with the other five segments for average + standard deviation features in the testing data is as follows:

- segment 01 with the other five segments: 0.9714

Table 4.5: The cosine similarity results for different segments with average + standard deviation features in the test data after data preprocessing and oversampling of low label data distribution.

	01	**02**	**03**	**04**	**05**	**06**
01	1.000	0.977	0.963	0.964	0.970	0.983
02		1.000	0.953	0.961	0.955	0.977
03			1.000	0.957	0.938	0.973
04				1.000	0.955	0.974
05					1.000	0.963
06						1.000

- segment 02 with the other five segments: 0.9642
- segment 03 with the other five segments: 0.9566
- segment 04 with the other five segments: 0.9624
- segment 05 with the other five segments: 0.9522
- segment 06 with the other five segments: 0.9742

To ensure consistency with our previous selection, we also chose the first and fourth segments in the test data, which included the standard deviation feature, for validating the proposed method. Table4.6 shows the results of the six segments after adding the SD features.

4.3.6 Normalization Method

Normalization[68] in deep learning is a critical preprocessing step that involves scaling numerical data to a standard range. This process improves the performance and stability of neural networks. It improves the convergence speed, prevents numerical instability, ensures equal contribution of features, and enhances the overall model accuracy. In this experiment, we applied the standardscaler[74] method to normalize the training and validation data. standardscaler is a data normalization technique provided by the scikit-learn library in Python. It is used to standardize the features of a dataset so that they have a mean of zero and a standard deviation of one, ensuring that the input features have a uniform scale and allowing models to perform better.

We applied different normalization methods to validate our proposed data processing method. The first approach was to normalize the entire training

Table 4.6: Precision, Recall, and F1 scores for different segments with average + standard deviation features in the testing data after data preprocessing.

Engagement	Acc	Low (Recall/Precision/F1)	High (Recall/Precision/F1)	Very High (Recall/Precision/F1)
segment 01 (Features+SD)	0.452	0.462/0.157/0.234	0.449/0.504/0.475	0.456/0.501/0.477
segment 02 (Features+SD)	0.471	0.218/0.126/0.160	0.546/0.508/0.526	0.419/0.497/0.454
segment 03 (Features+SD)	0.474	0.321/0.166/0.218	0.635/0.506/0.563	0.320/0.509/0.393
segment 04 (Features+SD)	0.462	0.308/0.113/0.166	0.644/0.509/0.568	0.286/0.534/0.373
segment 05 (Features+SD)	0.453	0.333/0.160/0.216	0.444/0.499/0.469	0.477/0.476/0.476
segment 06 (Features+SD)	0.442	0.247/0.144/0.182	0.565/0.482/0.520	0.333/0.448/0.382

and validation dataset as a whole. This is a commonly used normalization method in current machine learning research. It helps reduce bias between features in the training data, prevents numerical instability, and thereby improves model accuracy. However, due to the nature of time-series video data and individual variability, each video has its own unique data range. Therefore, we attempted to normalize each video individually to investigate whether the characteristics and personalization of individual videos have an impact on the experimental results. Thus, the second approach was to normalize each video sample individually. Additionally, we normalized the data on a per-second basis before merging them into the input data for the deep learning model.

In the experiment of Table 4.7, we selected the first segment for the non-oversampled part of the training, validation, and test data. The results indicate that the overall performance on the unbalanced data shows that normalizing the processed data as a whole and normalizing the input samples per second are superior to normalizing each video individually. Therefore, we ultimately chose the methods of normalizing the entire processed data and normalizing each second of the video.

Table 4.7: Precision, Recall, and F1 scores for different standard scaler normalization methods.

Engagement	Low (Recall/Precision/F1)	High (Recall/Precision/F1)	Very High (Recall/Precision/F1)
Entire train and val	0.806/0.702/0.751	0.474/0.525/0.498	0.539/0.544/0.541
Entire test	0.346/0.090/0.142	0.523/0.521/0.522	0.373/0.537/0.440
Each video train and val	0.386/0.342/0.363	0.401/0.407/0.404	0.352/0.384/0.367
Each video test	0.333/0.057/0.097	0.385/0.491/0.432	0.336/0.468/0.392
Per-second train and val	0.475/0.599/0.530	0.591/0.445/0.507	0.421/0.527/0.468
Per-second test	0.218/0.118/0.153	0.658/0.511/0.575	0.295/0.486/0.367

4.4 Experiment Setting

4.4.1 Training and Testing Data Pattern

In Experiment Part 1, to validate our proposed oversampling method, we selected the DAiSEE dataset for verification due to its more imbalanced yet larger data volume. The original video data in the DAiSEE dataset[24] were divided for training, validation, and testing purposes with proportions of 60%, 20%, and 20%, respectively. To better train the model, we first combined the training set and the validation set. During model training, we then split this combined set into an 80:20 ratio to ensure the model was thoroughly trained.

In Section 3.2.4, we discussed using combinations such as average, average+ standard deviation, and average+ standard deviation + minimum/maximum to validate our proposed method. In Section 4.3.4, we compared the cosine similarity of the 6 non-oversampled segment data and the average cosine similarity of each single segment with the other five segments. Based on this comparison, we chose the first and fourth segments for model validation. Therefore, in the experimental phase, we have the following data combinations.

The experimental data combinations used the first and fourth segments for testing.

- The training and validation data included the following combinations:

- The testing included the following combinations:

Table 4.8: Experimental data combinations for training and validation

Data Combinations for Training and Validation
LSTM (Key Points)
LSTM-FCN (Key Points)
LSTM (Features)
LSTM-FCN (Features)
LSTM (Features+SMOTE)
LSTM-FCN (Features+SMOTE)
LSTM (Features+RMA)
LSTM-FCN (Features+RMA)
LSTM (Features+SMA)
LSTM-FCN (Features+SMA)
LSTM (Features+SMA+OS)
LSTM-FCN (Features+SMA+OS)
Skipped Moving Average (oversampled)
Skipped Moving Average (oversampled) + Standard Deviation
Skipped Moving Average (oversampled) + Standard Deviation + Min/Max Values

4.4.2 LSTM Model and Experimental Parameters

We applied PyTorch to build our experiment models. PyTorch is an open-source machine learning library based on the Torch library. It is widely used for various applications, including computer vision and natural language processing, due to its flexibility and ease of use in building deep learning models[67]. PyTorch provides dynamic computational graphs, which allow for more intuitive model building and debugging, making it a preferred choice for many researchers and practitioners in the field of deep learning.

First, we employed the StandardScaler method to normalize the merged training and validation datasets, which consisted of 32-dimensional features over 10 timesteps. This step ensures that the features have a mean of zero and a standard deviation of one, which helps in stabilizing and accelerating the learning process of the model. Next, we set the random state to 10 to maintain consistency and reproducibility in our experiments. Then, the processed data were trained using the LSTM model introduced in Section 4.2.1.

The proposed models were trained over the course of 50, 100, and 200 epochs, with dropout rates ranging from 0 to 0.5, to ensure that the models could adequately learn and generalize from the training data. In our model, we used the Adam optimizer with a learning rate set to 0.001. Each training

Table 4.9: Experimental data combinations for testing

Data Combinations for Testing
LSTM (Key Points)
LSTM-FCN (Key Points)
LSTM (Features)
LSTM-FCN (Features)
LSTM (Features+SMOTE)
LSTM-FCN (Features+SMOTE)
LSTM (Features+RMA)
LSTM-FCN (Features+RMA)
LSTM (Features+SMA)
LSTM-FCN (Features+SMA)
LSTM (Features+SMA+OS)
LSTM-FCN (Features+SMA+OS)
Skipped Moving Average
Skipped Moving Average + Standard Deviation
Skipped Moving Average + Standard Deviation + Min/Max Values

duration was conducted five times for thorough evaluation. After training, we evaluated the models using the original test set to gauge their performance on unseen data. For testing data processing, we applied the StandardScaler normalization technique by applying the normalization rules learned from the training data to the test data. This ensures consistency between the training and testing phases. Additionally, we employed the Skipped Moving Average method without oversampling. This involved undersampling the data to obtain 60 sequences per video, and then selecting one sequence per second to create a 10-timestep input structure. Finally, the best-performing result from the tests was selected as the evaluation metric for the model.

4.4.3 LSTM-FCN Model and Experimental Parameters

Similar to the previous LSTM model, we used PyTorch to build our experiment models in the LSTM-FCN part. All parameters, except for the model-specific ones, remain consistent with the LSTM experiment.

In the model section, we trained the data using an LSTM-FCN model. The forward pass involves processing the input through an LSTM layer followed by three stacked temporal convolutional blocks with filter sizes of 128,

256, and 128, respectively. Each of these blocks comprises a temporal convolutional layer, batch normalization, and a ReLU activation function. The outputs from the LSTM and convolutional layers are concatenated and then passed through a fully connected layer with softmax activation functions to generate the final output, as illustrated in Figure4.3.

To maintain consistency and comparability in our experiments, we also used 50, 100, and 200 epochs with dropout rates ranging from 0 to 0.5 as training parameters. The testing phase employed the same data and model parameter settings as the LSTM model section. Then, the best-performing result from the tests was selected as the evaluation metric for the model. Similarly, the best-performing result from the tests was selected as the evaluation metric for the model.

4.5 Results for Oversampling in Class Imbalanced Datasets

In evaluating classification tasks in machine learning and deep learning, commonly used metrics include accuracy, precision, recall, F1 score, and the confusion matrix, which are employed to measure model performance in estimation and detection tasks. However, due to the severe class imbalance in the DAiSEE dataset, this study adopts the F1 score as the primary evaluation criterion, while accuracy, precision, and recall are used as supplementary metrics. Accuracy is appropriate for situations where the data is balanced but may not be suitable for imbalanced datasets[27]. Precision and recall often involve a trade-off: improving one may reduce the other due to differences in dataset distribution[77]. Therefore, the F1 score is more suitable for addressing class imbalance issues because it represents the harmonic mean of precision and recall, combining both strengths. It is particularly effective in scenarios where a balance between precision and recall is required, especially in datasets with significant class imbalance.

In this chapter, we will present the validation and test results of the various data processing methods mentioned earlier, along with the comparative results of different parameters.

- Tables 4.10 and 4.11 present the experimental results of the original-data without any processing, as well as the results after applying the skipped moving average and skipped moving average with oversampling methods. Additionally, the results are compared with the RMA and SMOTE oversampling method. The first segment of high-distribution

data that was not oversampled was selected as the training and validation data. The results shown in Tables 4.10 and 4.11 have been previously published in study[70].

Table 4.10: Validation results for the original data, RMA, SMOTE, and Skipped Moving Average with 32-D features[70]

.

Engagement	Low (Recall/Precision/F1)	High (Recall/Precision/F1)	Very High (Recall/Precision/F1)
LSTM (Key Points)	0.101/0.211/0.137	0.553/0.595/0.573	0.607/0.526/0.564
LSTM-FCN (Key Points)	0.049/0.211/0.079	0.654/0.562/0.605	0.500/0.553/0.525
LSTM (Original)	0.069/0.114/0.086	0.587/0.558/0.572	0.482/0.491/0.486
LSTM-FCN (Original)	0.049/0.211/0.079	0.654/0.562/0.605	0.500/0.553/0.525
LSTM (SMOTE)	0.821/0.751/0.784	0.539/0.557/0.548	0.556/0.579/0.567
LSTM-FCN (SMOTE)	0.792/0.690/0.738	0.538/0.530/0.534	0.515/0.598/0.554
LSTM (Features+RMA)	0.727/0.664/0.694	0.536/0.523/0.530	0.512/0.572/0.540
LSTM-FCN (Features+RMA)	0.629/0.588/0.608	0.508/0.510/0.509	0.522/0.549/0.535
LSTM (SMA)	0.096/0.235/0.137	0.634/0.561/0.595	0.521/0.558/0.539
LSTM-FCN (SMA)	0.036/0.348/0.065	0.694/0.547/0.612	0.502/0.590/0.543
LSTM (SMA+OS)	0.806/0.702/0.751	0.474/0.525/0.498	0.539/0.544/0.541
LSTM-FCN (SMA+OS)	0.637/0.623/0.630	0.527/0.498/0.512	0.510/0.557/0.533

Tables 4.10 and 4.11 show the validation and testing results of the proposed method compared to the RMA and SMOTE oversampling technique in our study. The "LSTM (Original)" and "LSTM-FCN (Original)" results refer to the original data with facial and body features, without applying the moving average.

"LSTM (SMA)" and "LSTM-FCN (SMA)" show the results of the input data processed by the Skipped Moving Average (SMA) method, which uses a moving average window of 30 frames over 10 timesteps. "LSTM (SMA+OS)" and "LSTM-FCN (SMA+OS)" present the outcomes of training data processed by both the Skipped Moving Average and oversampling methods.

Table 4.11: Testing results for the original data, RMA, SMOTE, and Skipped Moving Average with 32-D features[70].

Engagement	Low (Recall/Precision/F1)	High (Recall/Precision/F1)	Very High (Recall/Precision/F1)
LSTM (Key Points)	0.069/0.114/0.086	0.587/0.558/0.572	0.482/0.491/0.486
LSTM-FCN (Key Points)	0.014/0.111/0.025	0.694/0.518/0.594	0.300/0.434/0.355
LSTM (Features)	0.030/0.380/0.060	0.760/0.570/0.650	0.400/0.560/0.470
LSTM-FCN (Features)	0.056/0.190/0.086	0.597/0.557/0.576	0.477/0.479/0.478
LSTM (Features+SMOTE)	0.295/0.053/0.089	0.385/0.475/0.425	0.350/0.487/0.407
LSTM-FCN (Features+SMOTE)	0.179/0.037/0.061	0.579/0.503/0.539	0.241/0.558/0.336
LSTM (Features+RMA)	0.320/0.082/0.131	0.499/0.528/0.513	0.403/0.527/0.457
LSTM-FCN (Features+RMA)	0.267/0.073/0.115	0.529/0.520/0.525	0.384/0.539/0.448
LSTM (Features+SMA)	0.192/0.109/0.140	0.665/0.510/0.577	0.314/0.526/0.393
LSTM-FCN (Features+SMA)	0.038/0.071/0.050	0.728/0.526/0.611	0.355/0.553/0.433
LSTM (Features+SMA+OS)	0.346/0.090/0.142	0.523/0.521/0.522	0.373/0.537/0.440
LSTM-FCN (Features+SMA+OS)	0.269/0.063/0.103	0.561/0.512/0.535	0.312/0.562/0.401

To effectively compare with existing oversampling methods, we also applied the RMA and SMOTE techniques to oversample the training data. "LSTM (Features+RMA)", "LSTM-FCN (Features+RMA)", "LSTM (SMOTE)" and "LSTM-FCN (SMOTE)" present the outcomes following the application of SMOTE oversampling. After processing with the Skipped Moving Average, the data samples comprised 60 sequences, with six sequences per second. We extracted one sequence from each second, forming a structure of 10 timesteps. Subsequently, the data labeled as low engagement were oversampled six times using the RMA and SMOTE techniques to serve as training data for the LSTM and LSTM-FCN models. This ensured uniformity in the data structure between the two oversampling methods: Skipped Moving Average oversampling, RMA and SMOTE oversampling.

LSTM (Original) and LSTM-FCN (Original) process the time-series data without applying undersampling or Skipped Moving Average techniques. In contrast, LSTM (SMA) and LSTM-FCN (SMA) involve Skipped Moving Average processing on the original data, without oversampling. The results clearly indicate a significant disparity between the low-engagement label and other labels, suggesting that data imbalances negatively impact classification outcomes.

However, after implementing Skipped Moving Average and oversampling, the results for the low-engagement labels in both LSTM and LSTM-FCN models (LSTM (SMA+OS) and LSTM-FCN (SMA+OS)) showed improvement compared to the settings without oversampling. This demonstrates that our proposed Skipped Moving Average oversampling method effectively addresses the class imbalance issues in time-series data.

Based on the test results of the LSTM model in Tables 4.11, our proposed Skipped Moving Average (SMA) oversampling method outperformed the RMA and SMOTE oversampling methods for both low and high distribution data. This result demonstrates the superiority of our method in handling oversampling for time-series data. For the LSTM-FCN model, the test results indicate that the SMA method shows better performance than the RMA and SMOTE methods for the low and very high labels in various metrics. Although the high label performance with the SMA oversampling method was not as good as with the RMA and SMOTE methods, it was still relatively close. This indicates that our proposed method also has a certain level of stability and effectiveness.

In the previous experiment, we observed that the LSTM (SMA+OS) data processing achieved the best results for low-distribution label data, with Recall/Precision/F1 scores of 0.346/0.090/0.142. This result also outperformed the RMA and SMOTE oversampling method. The F1 score of around 0.142 is the best among the comparisons but is not sufficiently accurate. Estimating low engagement in practical applications is not sufficient. However, our proposed method effectively mitigates the negative impact of class imbalance in engagement estimation studies. Therefore, in the context of this study, our results are relatively good. Therefore, in the following experiments, to further improve the performance of our proposed data processing method, we will retain the LSTM model parameters and use this result as a benchmark for subsequent experiments.

- The first segment data of the experimental results for the skipped moving average values, Standard Deviation and Extreme Values features.

Table 4.12: The first segment data of the experimental results for the skipped moving average values, Standard Deviation, and Extreme Values features in training and validation.

Engagement	Low (Recall/Precision/F1)	High (Recall/Precision/F1)	Very High (Recall/Precision/F1)
SMA	0.806/0.702/0.751	0.474/0.525/0.498	0.539/0.544/0.541
SMA+SD	0.844/0.731/0.784	0.510/0.542/0.526	0.506/0.538/0.522
SMA+SD+Min/Max	0.820/0.790/0.805	0.542/0.559/0.550	0.561/0.559/0.560

Table 4.13: The first segment data of the experimental results for the skipped moving average values, Standard Deviation, and Extreme Values features in testing.

Engagement	Low (Recall/Precision/F1)	High (Recall/Precision/F1)	Very High (Recall/Precision/F1)
SMA	0.346/0.090/0.142	0.523/0.521/0.522	0.373/0.537/0.440
SMA+SD	0.462/0.157/0.234	0.449/0.504/0.475	0.456/0.501/0.477
SMA+SD+Min/Max	0.256/0.124/0.167	0.518/0.514/0.516	0.440/0.501/0.468

In this experiment, we used the first segment data from the high-distribution part of the dataset, which had not undergone oversampling, as the experimental input. For the training and validation parts of these three sets of data, we applied the following methods respectively: Skipped Moving Average Oversampling, Skipped Moving Average Oversampling + Standard Deviation, and Skipped Moving Average Oversampling + Standard Deviation and Extreme Values. For the corresponding test data, we processed the data with Skipped Moving Average, Skipped Moving Average + Standard Deviation, and Skipped Moving Average + Standard Deviation and Extreme Values, but without applying oversampling.

Tables 4.12 and 4.13 present the experimental results. The test results indicate that the combination of Skipped Moving Average Oversampling + Standard Deviation for training and validation produced the best outcomes. For the low label engagement estimation, it achieved Recall/Precision/F1 scores of 0.462/0.157/0.234, which is an improvement in the F1 score by 0.092 compared to the previous experiment's result of 0.346/0.090/0.142. Although the performance for the high label did not surpass the previous experiment, the overall average performance across the three metrics was the best, indicating a balanced result. Since the purpose of our experiment

is to address data imbalance in engagement estimation, the Skipped Moving Average Oversampling + Standard Deviation combination's experimental results perfectly demonstrate the effectiveness and superiority of our proposed method. The findings in this section are part of ongoing research and will be detailed in a future publication.

- The fourth segment data of the experimental results for the skipped moving average values, Standard Deviation and Extreme Values features.

Table 4.14: The fourth segment data of the experimental results for the skipped moving average values, Standard Deviation, and Extreme Values features in training and validation.

Engagement	**Low (Recall/Precision/F1)**	**High (Recall/Precision/F1)**	**Very High (Recall/Precision/F1)**
SMA	0.825/0.764/0.793	0.488/0.519/0.503	0.506/0.504/0.505
SMA+SD	0.806/0.696/0.747	0.537/0.538/0.538	0.473/0.542/0.505
SMA+SD+Min/Max	0.824/0.768/0.795	0.559/0.543/0.551	0.511/0.562/0.535

Table 4.15: The fourth segment data of the experimental results for the skipped moving average values, Standard Deviation, and Extreme Values features in testing.

Engagement	**Low (Recall/Precision/F1)**	**High (Recall/Precision/F1)**	**Very High (Recall/Precision/F1)**
SMA	0.282/0.106/0.154	0.546/0.511/0.528	0.385/0.512/0.439
SMA+SD	0.308/0.113/0.166	0.644/0.509/0.568	0.286/0.534/0.373
SMA+SD+Min/Max	0.256/0.105/0.149	0.496/0.495/0.495	0.408/0.482/0.442

In this part of the experiment, we used the fourth segment of data, which has a high distribution in the DAISEE dataset and has not undergone oversampling, as the experimental input. Similar to the previous experimental setup, we processed the training and validation parts of these three data groups using the following methods: Skipped Moving Average Oversampling, Skipped Moving Average Oversampling + Standard Deviation, and Skipped Moving Average Oversampling + Standard Deviation and Extreme Values. For the corresponding test data, we applied Skipped Moving Average, Skipped Moving Average + Standard Deviation, and Skipped Moving

Average + Standard Deviation and Extreme Values, but without oversampling.

In the test results presented in Table 4.14 and Table 4.15, Skipped Moving Average Oversampling + Standard Deviation yielded the best performance for the low data distribution, with Recall/Precision/F1 scores of 0.308/0.113/0.166. Notably, the F1 score improved by 0.024 compared to the LSTM (SMA+OS) experiment from earlier tests. Although the Recall value did not perform as well as in previous experiments, the overall test results for the low-distribution label data showed an improvement.

- Comparison of three standardScaler normalization methods

In this experiment, we applied the StandardScaler normalization technique using three different methods to preprocess the dataset and evaluate their impact on model performance.

The three methods are as follows:

1. Overall Normalization:

   This method involves applying StandardScaler to the entire merged training and validation dataset. The goal is to collectively normalize the data to ensure consistency across all features.

2. Per-Video Normalization:

   In this approach, StandardScaler is applied individually to each video sample. This method aims to capture and normalize variations within each video, making the features comparable across different video samples.

3. Per-Second Normalization:

   This method involves normalizing each second of video samples in the DAiSEE dataset individually. After normalization, the segments are merged to form the input data for the deep learning model. This method is designed to handle temporal variations within the video data more effectively.

Tables 4.16 and 4.17 present the validation and testing results for three different StandardScaler normalization methods. In Table 4.11, among the experiments comparing unprocessed original data, Skipped Moving Average (SMA), Skipped Moving Average with oversampling (SMA+OS) and SMOTE oversampling methods, the LSTM (SMA+OS) data processing achieved the best results for low-distribution label data, with Recall/Precision/F1

Table 4.16: Validation results under different normalization methods.

Engagement	Low (Recall/Precision/F1)	High (Recall/Precision/F1)	Very High (Recall/Precision/F1)
Overall SMA+OS	0.806/0.702/0.751	0.474/0.525/0.498	0.539/0.544/0.541
Per-Video SMA+OS	0.386/0.342/0.363	0.401/0.407/0.404	0.352/0.384/0.367
Per-Second SMA+OS	0.475/0.599/0.530	0.591/0.445/0.507	0.421/0.527/0.468
Per-Second SMA+OS+SD	0.659/0.608/0.632	0.419/0.516/0.463	0.600/0.525/0.560

Table 4.17: Testing results under different normalization methods.

Engagement	Low (Recall/Precision/F1)	High (Recall/Precision/F1)	Very High (Recall/Precision/F1)
Overall SMA+OS	0.346/0.090/0.142	0.523/0.521/0.522	0.373/0.537/0.440
Per-Video SMA+OS	0.333/0.057/0.097	0.385/0.491/0.432	0.336/0.468/0.392
Per-Second SMA+OS	0.218/0.118/0.153	0.658/0.511/0.575	0.295/0.486/0.367
Per-Second SMA+OS+SD	0.317/0.128/0.182	0.512/0.507/0.509	0.419/0.502/0.457

scores of 0.346/0.090/0.142. Therefore, we use this result as the comparison data sample for different normalization methods. This part of the experiment involves applying overall normalization to the processed data.

"Per-Video SMA+OS" refers to normalizing each video input sample separately and then merging the processed sample data for model input. "Per-Second SMA+OS" refers to normalizing each second of the video before combining them into validation and test data.

In Table 4.13, the first segment's validation and training data, which added the standard deviation feature, achieved the best results in both precision improvement for low-distribution data and the balance of the three classification data. Therefore, in this set of experiments, we also compare the test results after normalizing the data with the added standard deviation feature for each second.

From the experimental results, it is observed that the Per-Second Normalization method yields the best overall performance in terms of recall, precision, and F1 scores across different engagement labels, both in "Per-

Second SMA+OS"and "Per-Second SMA+OS+SD."Additionally, the performance of the data with added standard deviation surpasses that of the data processed only with oversampling. This indicates the sensitivity and importance of standard deviation in engagement estimation experiments for detecting the physical characteristics of online learning students. The results also suggest that handling temporal variations more granularly by normalizing per-second segments is more effective for improving model performance. However, compared to the results in Table 4.13, where SMA+SD (adding standard deviation but applying overall normalization to the data) showed superior performance in terms of Recall/Precision/F1, the results are still better than this set of experiments.

Chapter 5

Experimentation with Skipped Moving Average for Transfer Learning

5.1 Purpose

The results from Experiment 1 demonstrate that our proposed method is effective in addressing the issue of data imbalance. Additionally, it raises the question of whether our method can also adapt to the diversity and variability inherent in time-series data. Therefore, in this experimental phase, we aim to apply transfer learning techniques to verify the effectiveness of our proposed SMA method under different data conditions. Specifically, we will evaluate whether the SMA approach can successfully handle the unique challenges presented by varying time-series data.

5.2 Application of Transfer Learning

In the field of engagement estimation research, data imbalance, data insufficiency, and the limited number of computer vision-related features are major issues. Additionally, the time span of a single online learning session is relatively long, making the ten-second videos used in the DAiSEE dataset from Experiment 1 seem almost fleeting in the context of online learning. If student engagement is detected every ten seconds and reported to teachers and students, it could potentially disrupt the continuity of the lesson and interfere with the students' concentration. Online course instructors are more interested in capturing student engagement over a more extended period, so ten-second video inputs are too short for effective engagement estimation.

Therefore, it is worth investigating whether our proposed skipping moving average method is also applicable to longer time-span data.

Collecting a large number of video samples from online learning courses with accurate student engagement labels that meet our requirements is undoubtedly time-consuming and costly. Therefore, in this part of the experiment, we use transfer learning to assess whether our proposed method is effective in different situations. Transfer learning is a machine learning technique where a model trained on one task is repurposed and fine-tuned for a different but related task. This approach leverages the knowledge gained from the original task to improve the performance and efficiency of the new task, often requiring less data and computational resources compared to training a model from scratch. We use transfer learning, where a model is pre-trained on the large DAISEE dataset and then fine-tuned on a combination of the "in the wild" dataset and our own dataset.

- The structure of the transfer learning model

We constructed our transfer learning model on the Keras platform due to its user-friendly and intuitive API, which simplifies the process of building and training deep learning models. Keras is highly modular, enabling flexible combinations and customization of components, which is essential for research. It seamlessly integrates with TensorFlow, providing access to powerful computational features and pre-trained models for rapid development.

In the experiment setting, we initially trained a model on the DAiSEE dataset. To enhance interpretability and flexibility, we incorporated an inner model with one neural network layer into the pre-training model. This design allows us to better understand and manipulate the features learned during pre-training.

After the initial training phase, we extracted the trained inner model from the pre-trained model and froze its parameters. Freezing the inner model ensures that the valuable information it has learned from the DAiSEE dataset is preserved and not altered during subsequent training phases.

Next, we augmented the frozen inner model with a new trainable LSTM (Long Short-Term Memory) neural network layer. This LSTM layer is added at the bottom of the frozen layers, allowing it to leverage the pre-trained feature extraction capabilities while being specifically trained on the new dataset. By doing this, we ensure that the model can utilize the generalized features learned from the large DAiSEE dataset and adapt them to the specific characteristics of the new dataset.

Finally, we fine-tuned the newly built LSTM model. Fine-tuning involves making small adjustments to the model parameters to optimize performance

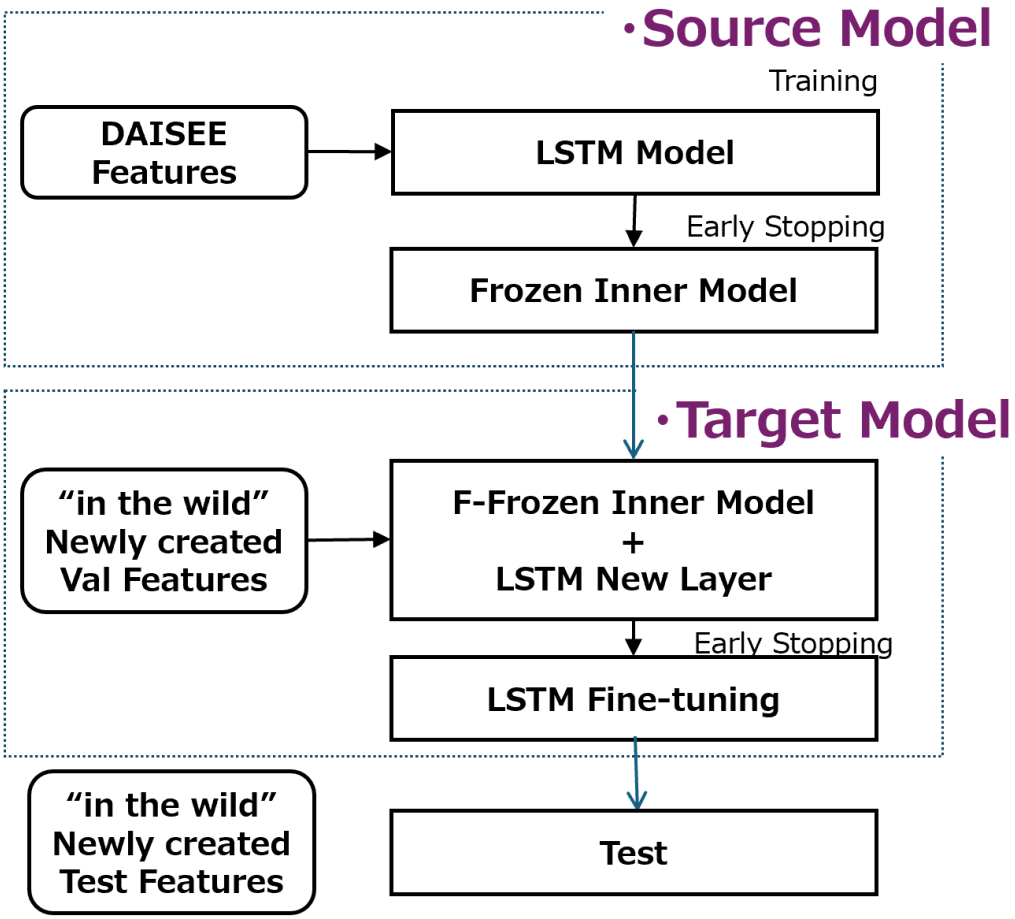

Figure 5.1: The structure of the transfer learning model applied in our study.

on the new dataset. This step ensures that the model not only retains the useful features learned from the DAiSEE dataset but also becomes highly effective for the specific task and data distribution of the new dataset. This process of integrating pre-trained models with new trainable layers and fine-tuning helps achieve better performance and robustness in the final model.

After fine-tuning the newly built model, it becomes our final model after transfer learning. To validate our fine-tuned model, we also tested it to ensure its performance. Figure 5.1 shows the structure of the transfer learning model we proposed.

5.3 Data preprocessing

In this set of experiments, we use three datasets. In transfer learning, we first need a relatively large dataset to train a model, and then transfer the trained model to a smaller dataset for fine-tuning. This approach improves the performance of the model trained on the smaller dataset. We use the DAiSEE dataset as the primary dataset, and then transfer the model trained on the DAiSEE dataset to the "in the wild" dataset and our own dataset for further training and fine-tuning. Therefore, in this section, we will introduce the data preprocessing methods for each of the three datasets separately.

- Data Preprocessing for the DAiSEE Dataset

In the transfer learning experiment, we also use the LSTM (SMA+OS) with Recall/Precision/F1 scores of 0.346/0.090/0.142 from Table 4.11 as a benchmark. Therefore, we apply the same preprocessing method to the DAiSEE dataset as before. Specifically, we extract body and facial keypoints using OpenPose and design features including eye information, eyebrow and lip shapes, facial rotation angles, head and body posture, the distance between the face and the screen, and body movements.

We then set a window period of 5 frames for the skipping moving average method and performed undersampling on the 10-second, 300-frame in the DAiSEE dataset. After undersampling, we obtain 60 segments. In the training and validation sets, to address the low distribution data, we oversample by keeping all data from the six segments labeled as low unchanged. For the high distribution labels, high and very high, we retain only the first segment as input samples for deep learning.

In the test data, no oversampling is performed on the samples from all labels to better reflect real online learning conditions. After preprocessing, the training and validation data for the low, high, and very high labels consist of the following numbers of samples: 2764, 4009, and 3286, respectively.

- Data Preprocessing for the "in the wild"Dataset

The "in the wild" dataset is also a popular public dataset for engagement estimation. This dataset includes 264 videos, each approximately five minutes long, with labeled engagement levels. The dataset comprises 91 subjects (27 females and 64 males) recorded in various settings such as computer labs, dorm rooms, and open spaces. In contrast, the data in the DAiSEE dataset is uniformly processed, with all videos being 10 seconds long, and recorded at 30 frames per second, resulting in 300 frames per video. Such a well-organized dataset is convenient for data processing and maintaining consistency.

Unlike the DAiSEE dataset, the "in the wild" dataset features videos of approximately five minutes in length, with significant variations in frame quality and count per video. Therefore, processing the "in the wild" dataset involves additional considerations:

1. Data Smoothing and Averaging:

   Ensuring smooth transitions and averaging data effectively.

2. Standardizing Video Lengths:

   Adjusting the irregular video lengths to create a uniform dataset suitable for deep learning model inputs.

Since the DAiSEE dataset consists of 10-second videos and the "in the wild" dataset contains five-minute videos, we need to maintain consistency in the data. Therefore, in Experiment 2, we segment the five-minute "in the wild" videos into 10-second segments for data preprocessing.

The method is as follows:

First, the five-minute video samples are segmented into 10-second videos to maintain an input of 10 timesteps per input video.

Next, body and facial features are calculated for the segmented 10-second video segments.

As shown in Figure 2.3, although the "in the wild" dataset also exhibits data imbalance, it is not as extreme as in the DAiSEE dataset. Therefore, when processing the "in the wild" data, we apply a skipping moving average without oversampling. Specifically, after segmenting the "in the wild" videos into 10-second segments, we process the 10-second videos with a moving window of 30 frames, averaging the values based on a frame rate of 30 frames per second. This ensures that each timestep of data corresponds to one second of input. With this processing, there is only one segment, and no segment selection is needed.

The original labels of the unsegmented videos are retained as the labels for the newly processed data.

However, this method presents an issue: due to the varying lengths of the videos, some segments may be shorter than 10 seconds. For video segments shorter than 10 seconds, we pad the insufficient part with values of -1 to ensure the consistency of the deep learning data input.

The same skipping moving window with a 30-frame interval is applied to the test data, and any segments shorter than 10 seconds are padded with a target value of -1.

The above content describes our method for processing the "in the wild" dataset.

- Data Preprocessing for the newly created Dataset

Unlike the previous two datasets, this new dataset involves answering 30 questions over approximately 12 minutes, resulting in significant variability in the length of each data segment. This increases the difficulty and complexity of data preprocessing. Therefore, we applied the following preprocessing steps to the dataset.

Firstly, the videos were clipped based on the time span of answering each question. Each clip corresponds to the duration a participant spent on a particular question. The engagement label provided by the participant's self-report and external observations from several study members, was assigned to the corresponding video clip. A comparison was made between the self-reports and the external observations, and in cases of disagreement, the final label was determined collectively by the observers. For each clipped portion, we extracted body and facial features, including eye information, eyebrow and lip shapes, facial rotation angles, head and body posture, the distance between the face and the screen, and body movements.

We applied a skipping moving average method with a window size based on the frame rate of 30 frames per second to smooth the data. This averaging window size ensures that each timestep corresponds to a consistent one second. Since the lengths of the clips varied, we ensured uniformity by padding clips shorter than the required length with a value of -1.

5.4 Experiment Setting

In this section, we will introduce the division of datasets and experimental parameters for Experiment 2.

5.4.1 Division and Usage of the Three Datasets

As described in Section 5.3, Data Preprocessing, we preprocessed the DAiSEE, "in the wild," and newly created datasets separately. The structure of the preprocessed data is shown in Table 5.1.

All the processed data from DAiSEE (SMA+OS) was retained as the primary dataset for transfer learning. The "in the wild" (SMA) and New Created (SMA) datasets were mixed at an 80:20 ratio to form the fine-tuning validation data and the final testing data for the fine-tuned model.

At the fine-tuning stage, we made the following data adjustments to enhance the performance of the fine-tuning process. One limitation of the "in the wild" (SMA) dataset is that it consists of a single ethnicity. To enhance

Table 5.1: Data distribution in transfer Learning Preprocessing.

Affective State	Very Low/Low	High	Very High
DAISEE(SMA+OS)	2764	4009	3286
"in the wild"(SMA)	985	1479	833
Newly Created(SMA)	392	610	360

the generalizability of the model, we added a portion of the newly created dataset to the "in the wild" (SMA) dataset based on the same 80:20 ratio. This addition was made to increase the diversity of features in the fine-tuning dataset. At the same time, we ensured that there was no overlap between the validation data and the testing data. The sample videos in the 'in the wild' dataset are approximately five minutes. To maintain consistency with the other two datasets, the proposed method estimates in 10-second increments, as done in DAiSEE. After splitting the five-minute video data into 10-second segments, each segment is evaluated using the original engagement label as its label.

This preprocessing approach helps maintain a balanced and diverse dataset for fine-tuning and testing, thereby improving the robustness and performance of the final transfer learning model.

5.4.2 Experimental Parameters

In our transfer learning experiments, we used distinct sets of parameters for training the model on the source dataset (DAiSEE) and for fine-tuning the model on the target datasets ("in the wild" and the newly created dataset). For training on the DAiSEE dataset, the parameters were: 500 epochs, 16 hidden units, a dropout rate of 0.1, a batch size of 32, a 'softmax' activation function, the 'adam' optimizer, and early stopping callbacks.

These parameters were chosen to ensure robust training and to prevent overfitting.

For fine-tuning the target datasets, we adjusted the parameters to ensure the model could adapt to the new data while preserving the learned features. The fine-tuning parameters included a learning rate of 0.00001, a 'softmax' activation function, 500 epochs, a batch size of 32, and early stopping callbacks.

5.5 Results for Transfer Learning

We used the "in the wild" and newly created datasets, split at an 80:20 ratio, to train an LSTM model as a baseline for comparison. The transfer learning method mentioned in Experiment 2 was then applied to validate whether our proposed data processing method positively impacts the model trained on data from different scenarios.

Table 5.2: Validation results of the model on "in the wild" and newly created datasets in LSTM model and transfer learning model.

Engagement	Acc.	**Low (Recall/Precision/F1)**	**High (Recall/Precision/F1)**	**Very High (Recall/Precision/F1)**
LSTM Model	0.892	0.892/0.931/0.911	0.909/0.875/0.892	0.865/0.881/0.873
Transfer Learning (without OS)	0.582	0.565/0.706/0.627	0.780/0.526/0.628	0.299/0.622/0.404
Transfer Learning (with OS)	0.578	0.659/0.667/0.663	0.661/0.542/0.595	0.364/0.538/0.434

Table 5.3: Testing results of the model on "in the wild" and newly created datasets in LSTM model and transfer learning model.

Engagement	Acc.	**Low (Recall/Precision/F1)**	**High (Recall/Precision/F1)**	**Very High (Recall/Precision/F1)**
LSTM Model	0.560	0.612/0.693/0.650	0.695/0.522/0.596	0.299/0.479/0.368
Transfer Learning (without OS)	0.586	0.600/0.718/0.654	0.746/0.533/0.622	0.325/0.568/0.413
Transfer Learning (with OS)	0.593	0.635/0.720/0.675	0.695/0.543/0.610	0.390/0.556/0.458

As shown in Table 5.3, our proposed data processing method SMA with oversampling TL (with OS), when applied in transfer learning, achieved the best results. The overall accuracy improved by 3.3%, with the F1 score for the Low engagement label increasing by 2.5%. Interestingly, in the transfer learning without oversampling (TL without OS) experiment, although the overall model accuracy increased by 2.6%, the F1 score for the Low engagement label showed almost no improvement due to the data distribution of the DAiSEE dataset. In contrast, the oversampled data in DAiSEE resulted in better model performance. This demonstrates that the Skipped Moving Average with oversampling data processing method effectively alleviates the

negative impact of data class imbalance on transfer learning techniques. Furthermore, it improved the overall accuracy of the model by 1% compared to the model without oversampling. The results presented in this section are planned to be published in future work.

Chapter 6

Discussion

6.1 Discussion of Comparative Experiments

In this study, we proposed a data preprocessing method to address imbalances in time-series video data. We also validated several aspects of feature selection and time-series data processing that required verification. In this chapter, we will discuss the findings and limitations of our experiments.

- The reliability of our proposed skipping moving average oversampling method.

The results in table 4.7 demonstrate the effectiveness of our proposed Skipped Moving Average (SMA) oversampling method in processing class-imbalanced video time-series data. Unlike SMOTE, which synthesizes new samples in the feature space around existing minority class samples, our method uses real data for oversampling, preserving the authenticity and continuity of the video time-series data.

SMOTE works by generating new instances from existing minority cases, essentially creating synthetic samples that do not exist in the original data[66]. While SMOTE is widely used in machine learning for balancing datasets, it has notable drawbacks. One significant issue is the blindness in neighbor selection, which can disrupt the continuity and natural progression of time-series data, as highlighted in related investigations[53]. This synthetic nature of SMOTE-generated data can potentially harm the temporal characteristics essential for video data analysis. In contrast, our SMA method maintains the inherent variability and continuity of the original video time-series data. By using actual data segments for oversampling, our approach ensures that the generated data are both realistic and contextually consistent. This authenticity is crucial for maintaining the integrity of time-dependent features.

Our experimental results show that the SMA method not only enhances the accuracy of the model but also demonstrates stability across different datasets and conditions. The real-data-based oversampling approach leads to better generalization and robustness, as it avoids the pitfalls associated with synthetic data generation. Thus, the SMA method is a reliable and effective solution for addressing class imbalances in video time-series data.

- The Skipped Moving Average (SMA) values, Standard Deviation, and Extreme Values features in our proposed method.

We applied the Skipped Moving Average for smoothing the existing data, and in addition, we incorporated features related to Standard Deviation and Maximum/Minimum values.

The Skipped Moving Average values are crucial for smoothing the time-series data. By averaging values over a specified window and skipping certain intervals, the SMA helps to reduce noise and capture the underlying trend in the data. The Standard Deviation feature measures the variability or dispersion of the data points from the mean. In the context of time-series video data, it helps quantify the fluctuations in engagement levels over time. High standard deviation indicates greater variability, which can be important for detecting changes in user engagement and distinguishing between different states of attention or interest. The Extreme Values feature captures the minimum and maximum values within a given segment of the time-series data. These extreme values provide valuable information for identifying the most significant changes in the student's body movements over a given period.

Tables 4.12 and 4.13 present the experimental results. The results indicate that the experiments incorporating Standard Deviation and Extreme Values outperformed those using only the Skipped Moving Average. The test results demonstrate that the combination of Skipped Moving Average Oversampling and Standard Deviation during validation and testing produced the most favorable outcomes. Specifically, for the low engagement label, this combination achieved Recall/Precision/F1 scores of 0.462/0.157/0.234. This marks a notable improvement in the F1 score, which increased by 0.092 compared to the previous experiment's results of 0.346/0.090/0.142.

While the performance for the high engagement label in this experiment did not surpass the results from the previous approach, the overall average performance across all three metrics (Recall, Precision, and F1 score) was the best among the tested methods. This indicates that the combination of Skipped Moving Average Oversampling and Standard Deviation not only improved the accuracy for the low engagement label but also achieved a more balanced performance across different engagement levels.

Given that the primary objective of our experiment is to address data imbalance in engagement estimation, the results achieved with the Skipped Moving Average Oversampling and Standard Deviation combination strongly validate the effectiveness and superiority of our proposed method. This approach ensures that the model can handle imbalanced data more effectively, leading to more accurate and reliable engagement predictions across various categories. The improvements observed, particularly in the F1 score for the low engagement label, underscore the importance of incorporating Standard Deviation into the oversampling strategy.

However, the results show that the data incorporating only the Standard Deviation performed better than the data with both Standard Deviation and Extreme Values. This might be because Standard Deviation is a measure of the frequency of a student's body movements within a given time frame, which has a more significant relationship with engagement during online learning. The Extreme Values, on the other hand, represent the maximum amplitude of a student's body movement within a given period. If a student has large but infrequent body movements, it could still indicate that the student is paying attention. Conversely, if a student has low amplitude but high-frequency movements, it might suggest lower engagement. However, the inclusion of Extreme Values could interfere with the model's ability to accurately assess engagement levels, potentially leading to misinterpretation of the engagement degree. Therefore, this part of the experiment suggests that the frequency of body movements plays a more crucial role in determining engagement than the amplitude of those movements.

- Three forms of video time-series data normalization.

Tables 4.16 and 4.17 present the experimental results for three different normalization approaches: normalizing the entire dataset, normalizing each video individually, and normalizing each second of video data.

From the experimental results, it is observed that the Per-Second Normalization method yields the best overall performance in terms of recall, precision, and F1 scores across different engagement labels, particularly in the "Per-Second SMA+OS"and "Per-Second SMA+OS+SD"configurations. Additionally, the performance of the data with added standard deviation surpasses that of the data processed solely with oversampling. This aligns with previous experimental findings, where the inclusion of the standard deviation feature consistently outperforms the use of a simple moving average alone.

The results also suggest that handling temporal variations more granularly by normalizing per-second segments is more effective for improving

model performance. This approach allows the model to capture finer details in the time-series data, leading to better engagement estimation. However, when compared to the results in Table 4.13, where SMA+SD (with standard deviation added but overall normalization applied to the data) showed superior performance in Recall, Precision, and F1 scores, the Per-Second Normalization method still demonstrates a more nuanced improvement in this set of experiments.

- Impact of different time segment data on the model

In this set of experiments, we compared the cosine similarity between six segments of data that had not undergone oversampling, as well as the results of comparing each segment with the other five segments'average values. Additionally, we compared the cosine similarity between segments processed with only the Skipped Moving Average and those processed with both the Skipped Moving Average and the added Standard Deviation.

In the data processed with Skipped Moving Average oversampling, we used the first and fourth segments in the later experiments. The average cosine similarity between the first segment and the other five segments was 0.9262, while for the fourth segments, it was 0.9290. Notably, the fourth segment's similarity of 0.9290 was the highest average cosine similarity among all groups. In the test data, the cosine similarity of the first and fourth segments with the other five segments were 0.9866 and 0.9824, respectively, with 0.9866 being the highest value. Consistent with the results in Tables 2.1 and 4.15, segments with higher cosine similarity achieved better model testing outcomes.

In the comparison of cosine similarity between different segments after adding the Standard Deviation, it turns out that the first and fourth segments had the highest cosine similarity with the other segments, both reaching 0.846. Additionally, when comparing the similarities among all segments, the first segment again achieved the highest similarity, with a value of 0.927. This indicates that, after adding the Standard Deviation, the first segment consistently had the highest similarity, both in direct comparisons and in the average similarity with other data.

From our subsequent experiments, it is evident that the Skipped Moving Average oversampling method with the addition of Standard Deviation produced the highest results for the "low" label so far, while also achieving the most balanced overall accuracy. The (Recall/Precision/F1) values for the three engagement labels were 0.462/0.157/0.234, 0.449/0.504/0.475, and 0.456/0.501/0.470, respectively. These results provide new insights, showing that there is a positive correlation between the cosine similarity among

different segment data and the model's training outcomes. Based on the current experimental findings, higher cosine similarity appears to lead to better model performance.

- Performance of the skipped moving average oversampling method on datasets with different time spans

In the transfer learning phase, the application of our proposed transfer learning model led to a significant improvement, with the F1 score increasing by nearly 0.25. Additionally, there was a substantial enhancement in the overall balance of the test results, leading to a more stable and robust performance, particularly in the "Low" and "Very High" categories, compared to the LSTM model. This improvement suggests that the transfer learning approach effectively enhances the model's ability to generalize across diverse datasets and varying conditions.

The results from the transfer learning model on both the validation and test sets showed consistent performance across various metrics, indicating that the model trained on the source dataset has successfully adapted to the different data distributions in the target dataset. This outcome demonstrates that our proposed Skipped Moving Average Oversampling method is highly adaptable to various time-series data scenarios. It also confirms the method's effectiveness and high stability, making it a reliable approach for handling diverse data with temporal dynamics.

6.2 Error Analysis

The main purpose of this study is to estimate low engagement levels among learners during online learning to provide appropriate support when students lose engagement. Therefore, we analyzed the misclassified videos between the labels of engagement.

For the misclassification of low engagement labels shown in 6.1, learners' facial and bodily expressions usually exhibit little to no change, but their gaze tends to wander. In other words, no noticeable feature changes can be captured from the video, and subtle gaze changes are also difficult to detect. As a result, identifying low engagement videos without explicit changes in facial or bodily expressions becomes particularly challenging.

For the misclassification of high engagement shown in 6.2, two scenarios exist: one where it is mistakenly classified as very high engagement, and another where it is misclassified as low engagement. The first scenario, where high engagement is misclassified as very high engagement, is similar to the

Figure 6.1: Misclassification of low engagement labels.

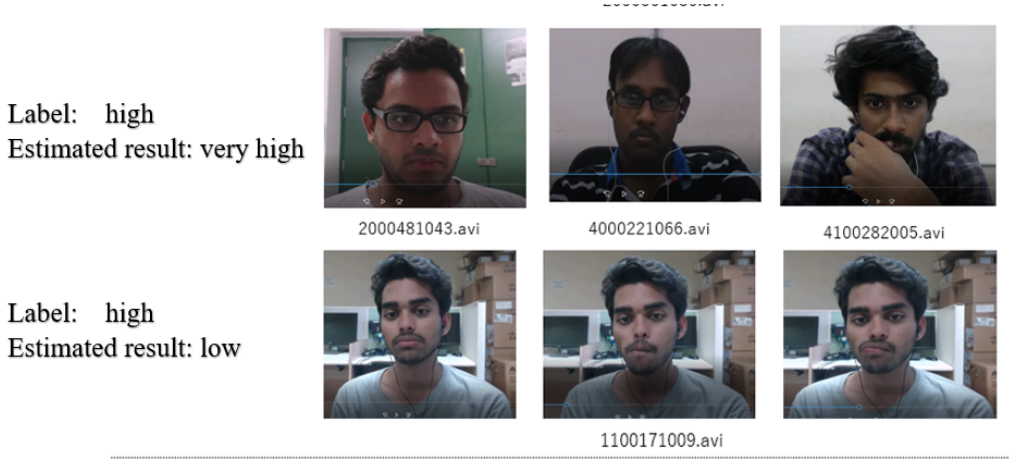

Figure 6.2: Misclassification of high engagement labels.

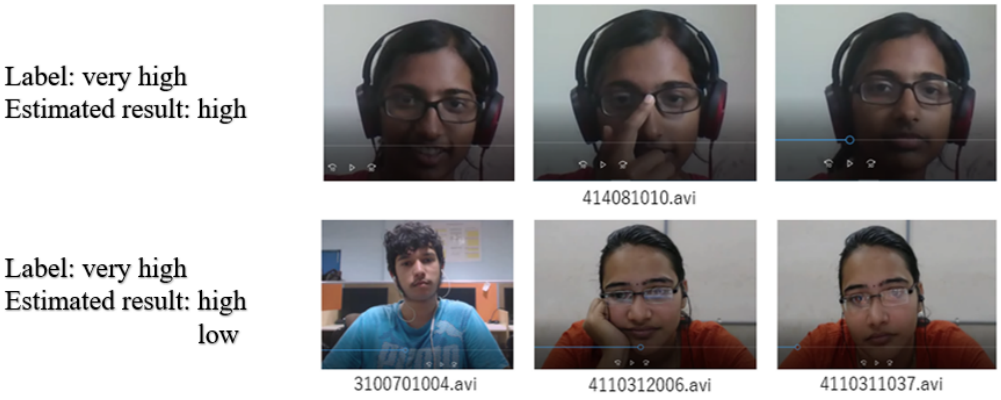

Figure 6.3: Misclassification of very high engagement labels.

misclassification of low engagement. In this case, facial and bodily expressions show minimal changes, but gaze variation is less than in low engagement misclassification. The second scenario, where high engagement is misclassified as low engagement, occurs when students listen attentively but exhibit behaviors like eating or drinking. These excessive physical movements introduce additional feature variations, which increase the difficulty for the classifier in distinguishing engagement levels accurately.

In cases where very high engagement is misclassified as shown in 6.3, it often occurs when students are attentively engaged in the online course content, but redundant facial and bodily feature information leads to misclassification into either high or low engagement levels. For instance, behaviors such as resting their hands on their faces, adjusting their glasses, eating, or drinking can result in such errors. In these videos, we observed that the learners'eyes and attention remained focused on the online course content; however, these extraneous bodily movements adversely affected the model's performance.

Another case involves the influence of individual characteristics, such as unique facial and bodily expressions or personal behavioral habits, which affect the model's performance. Personalized engagement estimation is an important research direction that warrants further attention.

From the analysis, the misclassification of videos labeled as low engagement is primarily due to insufficient information to accurately estimate learners' behavior. In the absence of clear facial and bodily expressions, capturing corresponding cues such as gaze direction, blinking, and eye movement becomes increasingly important. For the misclassification between high and very high engagement, the main issue lies in the excessive bodily feature in-

formation, which introduces potential interference with the model's performance. This suggests that, while bodily movements and actions enhance the performance of deep learning models, threshold adjustments are necessary to accommodate such anomalies. Additionally, the challenge of personalized engagement estimation is also significant. Improving the recognition of individual characteristics and behavioral habits is a crucial factor in ensuring the accuracy of foundational research. However, as observed in the analysis of misclassifications, the difference between high and very high engagement is often subtle and difficult to distinguish, even through external observation. This raises the question of how many labels are necessary for engagement estimation and classification studies, particularly for low engagement. Addressing this issue will be a vital topic for future research.

6.3 Gap Points for Improvement

From the above discussion, we can see that the proposed SMA method for processing video time-series data has shown some effectiveness in handling imbalanced time-series data. Furthermore, it has great potential in terms of both accuracy and stability when applied to video data of varying lengths and environments. However, there are still aspects of our experiments that need improvement.

The increase in accuracy observed after adding the SD feature highlights the importance of body vibration frequency in detecting learner engagement during online learning. We should further analyze and validate which body features have the greatest impact on our research. Additionally, further research and analysis should be conducted on features related to body movement.

In the comparative experiment of the three normalization methods, the method that normalized each second of video separately before merging them into the input data for the deep learning model showed the best performance in terms of accuracy. However, since the focus of our current experiment is to verify whether our proposed method is effective in addressing the issue of time-series data imbalance, we did not conduct further analysis of this aspect. Based on the results we have presented, normalization techniques could also serve as a breakthrough point for improving the performance of our future research.

In this experiment, we applied transfer learning to evaluate the performance of SMA on video data of different durations. However, in real online education environments, determining the most suitable duration for engagement research remains an important question. Furthermore, it is worth

exploring whether our proposed method can adapt to longer and more varied video time-series data. Another key research question for the future is whether the SMA method can be applied to time-series data prediction problems.

6.4 Advantages of the Skipped Moving Average Method

Our proposed method retains the characteristics of time series data while reducing its sensitivity to noise. The proposed skipped moving average oversampling method for time series data not only preserves the temporal characteristics of video data during oversampling but also addresses the issue of data imbalance in engagement estimation-related research. Moreover, it provides a reference approach for studies focused on estimating and predicting actions such as running, badminton, fitness, and other activities. Additionally, it reduces the sensitivity of time-series data to noise. In videos, the movement of learners inevitably results in inaccuracies in the features of the current frame, leading to errors in feature extraction. Such noise can negatively impact estimation results. However, the proposed SMA method smooths the curve of outliers between adjacent frames, thereby mitigating the adverse effects of noise on detection outcomes.

The proposed method not only retains the temporal characteristics of the data but also preserves its authenticity. Most existing data oversampling methods inevitably compromise the authenticity of the original video time series data. This poses a significant challenge for ensuring the effectiveness of the oversampled data in subsequent analysis. Furthermore, the similarity between adjacent levels of emotional engagement labels higher demands on input features. Our proposed method not only preserves the authenticity of the original data but also effectively redistributes the data across different classes. In addition, our method preserves the authenticity of video time series data while reducing computational cost and processing time. Although model-based video oversampling methods can retain data authenticity, their complexity and high generation cost present unavoidable challenges for real-time engagement estimation. In contrast, our proposed method not only retains data authenticity but also reduces the computational cost, ensuring support for future real-time engagement estimation and rapid processing tasks.

The proposed method addresses the issue of data imbalance to a certain extent. Although the current results indicate that there is still significant

space for improvement, it has achieved notable progress, providing an effective solution to the problem of imbalance in time series data. This method can be applied to various experimental approaches, such as time series model training and transfer learning. The generalizability of our proposed method is relatively high. It does not have specific requirements for the data and can be applied as long as it is video data. Moreover, the flexibility in selecting sliding windows enables its application to different estimation/detection targets and model tasks. Researchers can adjust the parameters according to the specific requirements of their studies and the data they need to process.

It provides feasibility for predicting engagement levels. One of the advantages of time series estimation/detection tasks is the ability to retain temporal dynamic information, enabling tasks such as action recognition and behavior prediction. Our ultimate goal is to predict changes in students' engagement levels during online learning and provide timely support and intervention when a student is predicted to lose engagement. This aims to help students achieve better learning outcomes during online education. Therefore, addressing the current challenges in engagement estimation/detection can also drive future research on engagement prediction.

Chapter 7

Conclusion

7.1 Summary and Contributions

In this study, we tackled the challenge of class-imbalanced time-series video data in the context of engagement estimation and detection by introducing a novel approach: Skipped Moving Average (SMA) oversampling. This method was specifically designed to address the complexities of video time-series data, where traditional oversampling techniques may fall short. By focusing on the unique temporal dynamics of video data, the Skipped Moving Average oversampling method enhances the accuracy and reliability of engagement level analysis. This approach not only mitigates the effects of class imbalance but also preserves the continuity and authenticity of the time-series data, leading to more precise and consistent results in engagement detection.

Throughout the overall experiment, we implemented the following procedures.

1. RQ1: How can we address the issue of class imbalance in datasets like DAiSEE?

- Facial and body Features

First, we used facial and body features for learners in online learning based on existing psychological research on the relationship between internal states and external expressions. The features include eye information, eyebrow and lip shapes, facial rotation angles, head and body posture, the distance between the face and the screen, and body movements. These features were selected to capture the subtle cues that reflect a learner's engagement and cognitive state during online learning sessions.

- Skipped Moving Average Oversampling

To address the issue of class imbalance in video time-series data, we proposed a novel data processing method known as the Skipped Moving Average (SMA) oversampling method. This technique was specifically developed to enhance the quality and balance of time-series data by selectively averaging data within specified intervals, effectively smoothing the data while preserving essential patterns. The SMA method ensures that the temporal characteristics of the video data are maintained, leading to more accurate and robust model performance. This method not only helps in balancing the dataset but also improves the model's ability to generalize across different time spans and scenarios, making it particularly effective for engagement estimation in diverse online learning environments.

2. RQ2: How does the proposed method influence the accuracy of engagement estimation?

- Comparison of Moving Average with Standard Deviation and Extreme Values

To increase the diversity of engagement estimation data, we compared the results of models tested with data processed using only the Skipped Moving Average method versus those with the addition of Standard Deviation and Extreme Values. The data with added Standard Deviation consistently achieved the best results. This comparison highlights the importance of incorporating statistical features like Standard Deviation, which capture variability and enhance the model's ability to recognize different levels of engagement.

- Comparison of Different Normalization Methods

We conducted a comparative experiment to evaluate the effectiveness of different normalization methods: normalizing the entire dataset, normalizing each video individually, and normalizing each second of video data. The results showed that overall dataset normalization outperformed per-second video normalization, while normalizing individual videos yielded the poorest results.

- Selecting Different Segment Data

Effectively selecting segments from a dataset that do not require oversampling is crucial to ensuring the effectiveness of our method. Experimental results indicate that segments with the highest similarity to other data tend

to produce better outcomes. These segment data likely contain representative patterns and features that are more consistent with the overall data distribution, thus enhancing the model's performance when they are used as a foundation for training without oversampling.

- Model Training

In our experiments, we employed both LSTM and LSTM-FCN models to evaluate the effectiveness of our proposed methods. The LSTM model consistently outperformed the LSTM-FCN model in testing results. Additionally, to assess the performance of our proposed method under different conditions, we implemented a transfer learning framework to test the data processed using our approach.

3. Q3: Can the proposed method with fine-tuning adapt to different video datasets?

We applied our proposed SMA method to video data from three datasets with different time spans and types. Then, we used a transfer learning model to verify whether SMA could improve the accuracy of data with varying time spans. The results demonstrated the feasibility and stability of our proposed method.

Based on the results of the above experiments, it is clear that our proposed method has made a positive impact on the study of engagement estimation. Additionally, it has helped to address the issue of class imbalance in time-series data to some extent.

7.2 Further work

The training and test data duration is an indispensable aspect that drives this research forward. Making full use of existing public datasets is also an important issue. Therefore, testing our proposed methods on new datasets in different scenarios and environments that more closely resemble real online education settings will be an important focus for future work.

The length of individual video samples within a dataset is also a critical factor to consider. Using one second unit Skipped Moving Average may be too short for real online education scenarios, as changes in learner engagement are unlikely to occur frequently within a single class session. Therefore, determining the appropriate window size for the Skipped Moving Average oversampling method is an important aspect to explore further. Additionally,

deciding how frequently to report engagement levels to both teachers and students is another significant research question that needs to be addressed.

In this experiment, we adopted a set of computer vision-based facial and body features. While we achieved some success in our experiments, understanding the relationship between each specific facial and body information and the learner's internal state remains an important area for further investigation. If we can identify which features are most effective at capturing the learner's internal state, there is significant potential to further improve the accuracy of our model through more targeted feature.

Furthermore, designing and facilitating learning support using the estimation results is a key direction for future research. Since the ultimate goal of detecting learner engagement in online learning is to improve the quality of education, future work should focus on creating an online educational support system that provides real-time engagement feedback to both learners and instructors. This system would enable learners to better manage and monitor their own progress, while also allowing instructors to adjust their teaching content and quality based on engagement feedback, leading to a more effective and personalized educational experience.

The current state of generative AI for video-to-video synthesis involves significant advancements in creating videos by learning temporal and spatial features. Therefore, using existing datasets as a foundation to generate similar videos with the same labels for data oversampling is also a potential direction for future development. However, ensuring that the generated videos meet our requirements is a critical prerequisite. Research has shown that engagement levels with similar labels often exhibit very similar facial and body information. Thus, in the generated videos, ensuring that they meet the current requirements with accurate label information is essential. Additionally, the associated human costs and the computational costs of complex models involved in this process are inevitable factors that need careful consideration.

Additionally, it is worth noting that during our experiments, we observed that the processing speed for tasks such as data processing and model training was nearly real-time. This is a significant improvement in addressing the issue of delayed feedback[54] in learner engagement. It also highlights the broader applicability of our proposed method.

Publications

International Journal

1. X. Zheng, S. Hasegawa, W. Gu, and K. Ota. Addressing class imbalances in video time-series data for estimation of learner engagement: "Over Sampling with Skipped Moving Average", *Education Sciences*, vol.14, no.6, pp.556 (2024), doi:10.3390/educsci14060556

International Conferences

2. X. Zheng, M. T. Tran, K. Ota, T. Unoki, and S. Hasegawa. Engagement Estimation using Time-series Facial and Body Features in an Unstable Dataset, *Proceedings of the 30th International Conference on Computers in Education (ICCE 2022)*, Kuala Lumpur, Malaysia, 28 November–2 December 2022, pp.89–94

3. X. Zheng, S. Hasegawa, M. T. Tran, K. Ota, and T. Unoki. Estimation of learners'engagement using face and body features by transfer learning, *Proceedings of the International Conference on Human–Computer Interaction*, Virtual, 24–29 July 2021, Springer International Publishing, Cham, Switzerland, pp.541–552

4. S. Hasegawa, A. Hirako, X. Zheng, S. N. Karimah, K. Ota, and T. Unoki. Learner's mental state estimation with PC built-in camera, *Learning and Collaboration Technologies. Human and Technology Ecosystems: Proceedings of the 7th International Conference, LCT 2020, Held as Part of the 22nd HCI International Conference, HCII 2020*, Copenhagen, Denmark, 19–24 July 2020, Springer International Publishing, Berlin/Heidelberg, Germany, pp.165–175

International Workshops

5. X. Zheng. A Learner Engagement Estimation and Support System Using PC Built-in Camera, *Student session on International Workshop on Affective Interaction between Humans and Machines in Multicultural Society* (Presentation only), (2022)

6. X. Zheng, S. Hasegawa, K. Ota, and T. Unoki. Engagement estimation using facial and body information, in *ASEAN Workshop on Information Science and Technology (AWIST 2020)*

Domestic Conferences

7. 長谷川忍, X. Zheng, 太田光一, 卯木輝彦. ” 主体的学習における表情からのエンゲージメント推定のためのデータセット構築に関する検討”, 第 *44* 回教育システム情報学会全国大会講演論文集, pp.187–188 (2019)

Acknowledgments

My supervisor, Professor Hasegawa Shinobu, has provided me with much academic guidance throughout my journey from research student studies to the completion of my master's and doctoral degrees. He helped me with my research and coursework. He is a kind and warm person who always inspires hope and motivates others to move forward. During these seven years together, we all came to realize that Professor Hasegawa is actually a strict mentor. He is considerate of our feelings, so he only steps in at critical moments to remind us in his own way to work harder. I am especially grateful for his patience and support during the many times when I felt like giving up. His guidance and encouragement helped me push through and keep going. Meeting Professor Hasegawa and experiencing this journey with him has been one of the most beautiful and meaningful parts of my life.

My minor research supervisor, Professor Okada, provided me with invaluable advice during my master's thesis defense. He also gave me a great deal of freedom to explore areas of my interest during my doctoral minor research, which greatly contributed to the progress of my main research.

Ota-sensei, our lab assistant professor, has been a great source of support throughout my time at JAIST. I would like to express my gratitude for his care and for accompanying me on a business trip to Malaysia during the COVID-19 period to ensure my safety. Although he joined JAIST about a year after I did, he has consistently provided me with encouragement and comfort, especially during moments of stress. His optimism has been a source of strength for me.

I would like to express gratitude to Professor Okada, Professor Ikeda, Professor Kashihara, and Professor Xie for their invaluable questions and comments during my preliminary defense. Professor Okada and Professor Kashihara, through seemingly simple yet profound questions, guided me to deeply reflect on and recognize the shortcomings in my research and dissertation. Their approach provided me with the space to think independently and the opportunity to proactively improve my future research. Meanwhile, the suggestions from Professor Ikeda and Professor Xie offered me a new perspective to validate and ensure the effectiveness of our proposed method. Their insights have allowed me to gain a deeper understanding of my thinking processes.

I would like to thank my parents for their support throughout my studies. The absence of our time together over these past six years has left many regrets in our lives. However, their understanding and support have allowed me to persevere and complete my doctoral program.

Bibliography

[1] J. L. Moore, C. Dickson-Deane, and K. Galyen. e-Learning, online learning, and distance learning environments: Are they the same?, *The Internet and Higher Education*, vol.14, no.2, pp.129–135 (2011)

[2] A. Benson. Using online learning to meet workforce demand: A case study of stakeholder influence, *Quarterly Review of Distance Education*, vol.3, no.4, pp.443–452 (2002)

[3] D. Conrad. Deep in the hearts of learners: Insights into the nature of online community, *Journal of Distance Education*, vol.17, no.1, pp.1–19 (2002)

[4] S. R. Hiltz and M. Turoff. Education goes digital: The evolution of online learning and the revolution in higher education, *Communications of the ACM*, vol.48, no.10, pp.59–64 (2005), doi:10.1145/1089107.1089139

[5] M. D. B. Castro and G. M. Tumibay. A literature review: efficacy of online learning courses for higher education institution using meta-analysis, *Educ Inf Technol*, vol.26, pp.1367–1385 (2021), doi:10.1007/s10639-019-10027-z

[6] P. C. D. Oliveira, C. J. C. D. A. Cunha, and M. K. Nakayama. Learning Management Systems (LMS) and e-learning management: an integrative review and research agenda, *JISTEM-Journal of Information Systems and Technology Management*, vol.13, no.2, pp.157–180 (2016)

[7] B. S. Bell and J. E. Federman. E-learning in postsecondary education, *The Future of Children*, pp.165–185 (2013)

[8] E. Hargittai. The digital reproduction of inequality, *The Inequality Reader*, Routledge, pp.660–670 (2018)

[9] S. Z. Salas-Pilco, Y. Yang, and Z. Zhang. Student engagement in online learning in Latin American higher education during the COVID-19 pandemic: A systematic review, *British Journal of Educational Technology*, vol.53, no.3, pp.593–619 (2022)

[10] R. M. Ryan and E. L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being, *American Psychologist*, vol.55, no.1, pp.68–78 (2000)

[11] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris. School engagement: Potential of the concept, state of the evidence, *Review of Educational Research*, vol.74, no.1, pp.59–109 (2004)

[12] M. Hartnett. The importance of motivation in online learning, *Motivation in Online Education*, pp.5–32 (2016)

[13] E. R. Kahu. Framing student engagement in higher education, *Studies in Higher Education*, vol.38, no.5, pp.758–773 (2013)

[14] J. B. Arbaugh. Virtual classroom characteristics and student satisfaction with internet-based MBA courses, *Journal of Management Education*, vol.24, no.1, pp.32–54 (2000)

[15] D. R. Garrison and M. Cleveland-Innes. Facilitating cognitive presence in online learning: Interaction is not enough, *The American Journal of Distance Education*, vol.19, no.3, pp.133–148 (2005)

[16] M. Kage. *Theory of Motivation to Learn: Motivational Educational Psychology*, Kaneko Bookstore, (2013)

[17] M. A. A. Dewan, M. Murshed, and F. Lin. Engagement detection in online learning: A review, *Smart Learning Environments*, vol.6, pp.1 (2019)

[18] B. Hollister, P. Nair, S. Hill-Lindsay, and L. Chukoskie. Engagement in online learning: Student attitudes and behavior during COVID-19, *Frontiers in Education*, vol.7, pp.851019 (2022)

[19] F. Martin and D. U. Bolliger. Engagement matters: Student perceptions on the importance of engagement strategies in the online learning environment, *Online Learning*, vol.22, pp.205–222 (2018)

[20] J. Nouri. The flipped classroom: For active, effective and increased learning–especially for low achievers, *International Journal of Educational Technology in Higher Education*, vol.13, pp.1–10 (2016)

[21] D. U. Bolliger. Key factors for determining student satisfaction in online courses, *International Journal of E-Learning*, vol.3, pp.61–67 (2004)

[22] S. N. Karimah and S. Hasegawa. Automatic engagement estimation in smart education/learning settings: A systematic review of engagement definitions, datasets, and methods, *Smart Learning Environments*, vol.9, pp.1–48 (2022)

[23] A. Kaur, A. Mustafa, L. Mehta, and A. Dhall. Prediction and localization of student engagement in the wild, *Proceedings of the 2018 Digital Image Computing: Techniques and Applications (DICTA)*, Canberra, Australia, 10–13 December 2018, pp.1–8

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

[25] R. L. Allen and A. S. Davis. Hawthorne Effect, *Encyclopedia of Child Behavior and Development*, S. Goldstein and J. A. Naglieri, Eds., Springer, Boston, MA, USA, (2011)

[26] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study, *Intelligent Data Analysis*, vol.6, pp.429–449 (2002)

[27] D. Dresvyanskiy, W. Minker, and A. Karpov. Deep learning based engagement recognition in highly imbalanced data, *Proceedings of the 23rd International Conference, SPECOM 2021*, St. Petersburg, Russia, 27–30 September 2021, Springer International Publishing, Berlin/Heidelberg, Germany, pp.166–178 (2021)

[28] A. R. Anderson, S. L. Christenson, M. F. Sinclair, and C. A. Lehr. Check Connect: The importance of relationships for promoting engagement with school, *Journal of School Psychology*, vol.42, pp.95–113 (2004)

[29] A. L. Reschly and S. L. Christenson. *Handbook of Research on Student Engagement*, Springer Nature, Berlin/Heidelberg, Germany, (2022)

[30] M. Cocea and S. Weibelzahl. Log file analysis for disengagement detection in e-Learning environments, *User Modeling and User-Adapted Interaction*, vol.19, pp.341–385 (2009)

[31] N. R. Aljohani, A. Fayoumi, and S.-U. Hassan. Predicting At-Risk Students Using Clickstream Data in the Virtual Learning Environment, *Sustainability*, vol.11, no.24, pp.7238 (2019), doi:10.3390/su11247238

[32] Y. Liu, S. Fan, S. Xu, A. Sajjanhar, S. Yeom, and Y. Wei. Predicting Student Performance Using Clickstream Data and Machine Learning, *Education Sciences*, vol.13, no.1, pp.17 (2023), doi:10.3390/educsci13010017

[33] J. A. Fredricks. The Measurement of Student Engagement: Methodological Advances and Comparison of New Self-report Instruments, In: A. L. Reschly and S. L. Christenson (eds) *Handbook of Research on Student Engagement*, Springer, Cham, (2022)

[34] M. Chaouachi, C. Pierre, I. Jraidi, and C. Frasson. Affect and mental engagement: Towards adaptability for intelligent, *Proceedings of the Twenty-Third International FLAIRS Conference*, Daytona Beach, FL, USA, 19–21 May 2010

[35] S. H. Fairclough and L. Venables. Prediction of subjective states from psychophysiology: A multivariate approach, *Biological Psychology*, vol.71, pp.100–110 (2006)

[36] B. S. Goldberg, R. A. Sottilare, K. W. Brawner, and H. K. Holden. Predicting learner engagement during well-defined and ill-defined computer-based intercultural interactions, *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction, ACII 2011*, Memphis, TN, USA, 9–12 October 2011, Springer, Berlin/Heidelberg, Germany, pp.538–547

[37] Z. Zhang, Z. Li, H. Liu, T. Cao, and S. Liu. Data-driven online learning engagement detection via facial expression and mouse behavior recognition technology, *Journal of Educational Computing Research*, vol.58, pp.63–86 (2020)

[38] W. T. James. A study of the expression of bodily posture, *Journal of General Psychology*, vol.7, pp.405–437 (1932)

[39] A. Kleinsmith and N. Bianchi-Berthouze. Affective body expression perception and recognition: A survey, *IEEE Transactions on Affective Computing*, vol.4, pp.15–33 (2012)

[40] P. Ekman and W. V. Friesen. Measuring facial movement, *Environmental Psychology and Nonverbal Behavior*, vol.1, pp.56–75 (1976)

[41] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.24, pp.971–987 (2002)

[42] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection, *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, 20–26 June 2005, vol.1, pp.886–893

[43] C. Chang, C. Zhang, L. Chen, and Y. Liu. An ensemble model using face and body tracking for engagement detection, *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, Boulder, CO, USA, 16–20 October 2018, pp.616–622

[44] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.43, pp.172–186 (2021)

[45] J. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. Lester. Automatically recognizing facial expression: Predicting engagement and frustration, *Proceedings of the Educational Data Mining*, Memphis, TN, USA, 6–8 July 2013

[46] C. Chang, C. Zhang, L. Chen, and Y. Liu. An ensemble model using face and body tracking for engagement detection, *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, Boulder, CO, USA, 16–20 October 2018, pp.616–622

[47] S. M. Villaroya, J. J. Gamboa-Montero, A. Bernardino, M. Maroto-Gómez, J. C. Castillo, and M. Á. Salichs. Real-time engagement detection from facial features, *Proceedings of the 2022 IEEE International Conference on Development and Learning (ICDL)*, London, UK, 12–15 September 2022, pp.231–237

[48] X. Zheng, S. Hasegawa, M. T. Tran, K. Ota, and T. Unoki. Estimation of learners'engagement using face and body features by transfer learning, *Proceedings of the International Conference on Human–Computer Interaction*, Virtual, 24–29 July 2021, Springer International Publishing, Cham, Switzerland, pp.541–552

[49] X. Zheng, M. T. Tran, K. Ota, T. Unoki, and S. Hasegawa. Engagement Estimation using Time-series Facial and Body Features in an Unstable Dataset, *Proceedings of the 30th International Conference on Computers in Education (ICCE 2022)*, Kuala Lumpur, Malaysia, 28 November–2 December 2022, pp.89–94

[50] X. Ai, V. S. Sheng, C. Li, and Z. Cui. Class-attention video transformer for engagement intensity prediction, *arXiv*, arXiv:2208.07216 (2022)

[51] R. Mohammed, J. Rawashdeh, and M. Abdullah. Machine learning with oversampling and undersampling techniques: Overview study and experimental results, *Proceedings of the 2020 11th International Conference on Information and Communication Systems (ICICS)*, Irbid, Jordan, 7–9 April 2020, IEEE, Piscataway, NJ, USA, pp.243–248

[52] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, vol.16, pp.321–357 (2002)

[53] Z. Jiang, T. Pan, C. Zhang, and J. Yang. A new oversampling method based on the classification contribution degree, *Symmetry*, vol.13, pp.194 (2021)

[54] B. Yao, K. Ota, A. Kashihara, T. Unoki, and S. Hasegawa. Development of a Learning Companion Robot with Adaptive Engagement Enhancement, *Proceedings of the 30th International Conference on Computers in Education (ICCE 2022)*, Asia-Pacific Society for Computers in Education, Kuala Lumpur, Malaysia, 28 November–2 December 2022, pp.111–117

[55] M. A. A. Dewan, F. Lin, D. Wen, M. Murshed, and Z. Uddin. A deep learning approach to detecting engagement of online learners, *Proceedings of the 2018 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Guangzhou, China, 8–12 October 2018, pp.1895–1902

[56] M. Murshed, M. A. A. Dewan, F. Lin, and D. Wen. Engagement detection in e-learning environments using convolutional neural networks, *Proceedings of the 2019 IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, Fukuoka, Japan, 5–8 August 2019, pp.80–86

[57] N. Bosch. Detecting student engagement: Human versus machine, *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, Halifax, NS, Canada, 13–17 July 2016, pp.317–320

[58] M. Bond, K. Buntins, S. Bedenlier, O. Zawacki-Richter, and M. Kerres. Mapping research in student engagement and educational technology in higher education: A systematic evidence map, *International Journal of Educational Technology in Higher Education*, vol.17, pp.1–30 (2020)

[59] A. Mehrabian and J. T. Friar. Encoding of attitude by a seated communicator via posture and position cues, *Journal of Consulting and Clinical Psychology*, vol.33, pp.330–336 (1969)

[60] N. Dael, M. Mortillaro, and K. R. Scherer. Emotion expression in body action and posture, *Emotion*, vol.12, pp.1085–1101 (2012)

[61] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.43, pp.172–186 (2021)

[62] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 21–26 July 2017, pp.1145–1153

[63] P. Ekman. Facial expressions, In *Handbook of Cognition and Emotion*, John Wiley Sons, Ltd., Hoboken, NJ, USA, 1999, vol.16, p.e320

[64] S. Hochreiter and J. Schmidhuber. Long short-term memory, *Neural Computation*, vol.9, pp.1735–1780 (1997)

[65] F. Karim, S. Majumdar, H. Darabi, and S. Chen. LSTM fully convolutional networks for time series classification, *IEEE Access*, vol.6, pp.1662–1669 (2017)

[66] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, vol.16, pp.321–357 (2002)

[67] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, and L. Antiga. PyTorch: An imperative style, high-performance deep learning library, *Advances in Neural Information Processing Systems*, vol.32, MIT Press, Cambridge, MA, USA, 2019

[68] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization, *arXiv preprint*, arXiv:1607.06450 (2016)

[69] O. B. Adedoyin and E. Soykan. Covid-19 pandemic and online learning: The challenges and opportunities, *Interactive Learning Environments*, vol.31, no.2, pp.863–875 (2023)

[70] X. Zheng, S. Hasegawa, W. Gu, and K. Ota. Addressing class imbalances in video time-series data for estimation of learner engagement: "Over Sampling with Skipped Moving Average", *Education Sciences*, vol.14, no.6, pp.556 (2024), doi:10.3390/educsci14060556

[71] L. A. Jeni, J. F. Cohn, and F. De La Torre. Facing imbalanced data–recommendations for the use of performance metrics, *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, Geneva, Switzerland, 2–5 September 2013, IEEE, Piscataway, NJ, USA, pp.245–251

[72] S. Hasegawa, A. Hirako, X. Zheng, S. N. Karimah, K. Ota, and T. Unoki. Learner's mental state estimation with PC built-in camera, *Learning and Collaboration Technologies. Human and Technology Ecosystems: Proceedings of the 7th International Conference, LCT 2020, Held as Part of the 22nd HCI International Conference, HCII 2020*, Copenhagen, Denmark, 19–24 July 2020, Springer International Publishing, Berlin/Heidelberg, Germany, pp.165–175

[73] T. Baltrušaitis, P. Robinson, and L. P. Morency. OpenFace: An open source facial behavior analysis toolkit, *Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp.1–10, IEEE

[74] scikit-learn. (2023). StandardScaler —scikit-learn 1.5.1 documentation. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[75] DAiSEE Dataset for Affective States in E-Environments. Available online: https://people.iith.ac.in/vineethnb/resources/daisee/index.html (accessed on 31 July 2024)

[76] Seventh Emotion Recognition in the Wild Challenge (EmotiW). Available online: https://sites.google.com/view/emotiw2019/home?authuser=0 (accessed on 31 July 2024)

[77] Williams, C. K. The effect of class imbalance on precision-recall curves. *Neural Computation* **2021**, *33*(4), 853–857.

[78] Suresha, M.; Kuppa, S.; Raghukumar, D. S. A study on deep learning spatiotemporal models and feature extraction techniques for video understanding. *International Journal of Multimedia Information Retrieval* **2020**, *9*(2), 81–101.

[79] Islam, S.; Elmekki, H.; Elsebai, A.; Bentahar, J.; Drawel, N.; Rjoub, G.; Pedrycz, W. A comprehensive survey on applications of transformers for deep learning tasks. *Expert Systems with Applications* **2024**, *241*, 122666.

[80] Wu, W.; Hu, S.; Xiao, P.; Deng, S.; Li, Y.; Chen, Y.; Li, K. Video quality assessment based on swin transformer with spatio-temporal feature fusion and data augmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1846–1854, 2023.

[81] Wen, Q.; Sun, L.; Yang, F.; Song, X.; Gao, J.; Wang, X.; Xu, H. Time series data augmentation for deep learning: A survey. arXiv preprint arXiv:2002.12478, 2020.

[82] Gao, Z.; Liu, H.; Li, L. Data augmentation for time-series classification: An extensive empirical study and comprehensive survey. arXiv preprint arXiv:2310.10060, 2023.

[83] Iwana, B. K.; Uchida, S. An empirical survey of data augmentation for time series classification with neural networks. *PLoS ONE* **2021**, *16*(7), e0254841.

[84] Karimah, S. N.; Hasegawa, S. Automatic engagement recognition for distance learning systems: A literature study of engagement datasets and methods. In Proceedings of the International Conference on Human-Computer Interaction, pp. 264–276, Cham: Springer International Publishing, July 2021.

[85] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao. Time series classification using multi-channels deep convolutional neural networks, *Proceedings of the International Conference on Web-Age Information Management (WAIM)*, Cham, Switzerland, 16–18 June 2014, pp. 298–310, Springer International Publishing.