SPATIO-TEMPORAL REASONING FOR ESTIMATING STUDENT'S LEARNING AFFECT: AN APPROACH TO STRENGTHEN E-LEARNING

By

Christine Bukola Asaju

(1990591)

Dissertation
Submitted in fulfilment for the degree of
Philosophiae Doctor
in
Computer Science

School of Computer Science and Applied Mathematics
Faculty of Science
University of the Witwatersrand, Johannesburg
South Africa

Supervised by

Professor Hima B. Vadapalli

May 31, 2023

Abstract

Spatio-temporal reasoning has acquired traction in recent decades owing to its use in diverse application domains. It is an aspect of reasoning that involves drawing logical conclusions about changes in objects or entities over space and time. Reasoning technology has potential in various aspects of computer science, and the education sector is no exception. In a conventional classroom, a teacher reasons and canvases a student's comprehension by analyzing exhibited cues (via body language, speech, eye gaze, and facial emotions) and text-based feedback questions & answers (Q&A) to estimate the learning affect experienced. However, the limited existing work does not explore the usage of a combination of the above modalities in estimating comprehension. To date, researchers have focused on using facial cues exhibited by students to estimate learning affect, as existing research emphasizes that facial cues hold and convey relevant and meaningful information concerning a student's experienced learning affect. We believe that supplementing the above framework by incorporating reasoning based communication through question and answers has the potential to enhance the existing e-learning platforms and assist both teachers and students.

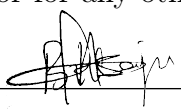
We propose a framework that includes affect analysis through facial emotion and affective state estimation, mapping of above estimates to learning affect as experienced by students, report generation for both students and teachers outlining the analysis by the model on students' comprehension of the lecture session and further feedback through natural text-based question and answering module. The proposed framework includes a Convolutional Neural Network-Bidirectional Long Short-Term Memory (CNN-BiLSTM) cascade to extract facial features and classify basic facial emotions and affective states. CNN-BiLSTM cascade handles and analyses video streams that capture a student's facial movements (temporal in nature) during a lecture session. The emotion estimator trained and tested on samples from Extended Denver Intensity of Spontaneous Facial Action (DISFA+) dataset annotated with seven basic emotions (i.e., anger, disgust, fear, happy, neutral, surprised and sadness) reported an accuracy of 92% and an average F1 score of 88% on a sample size of 2274. The mapping module then mapped the estimated emotions onto learning affects such as positive, negative, and neutral, based on mappings found in the literature. In similar lines, the affective state estimator trained and tested on the Dataset for the Affective States in E-Environment (DAiSEE) annotated with four affective states (i.e., boredom, confusion, engagement and frustration) reported an accuracy of 86% and an average F1 score of 87% on a sample size of 4,305. The aim of the affective state estimator was to classify emotion samples from DISFA+ and to use the existing affective state to learning affect mappings in literature to verify and confirm

the proposed emotion to learning affect mappings. The next stage of the proposed framework involves narrative report generation and feedback via questions and answers (Q&A). The narrative report generation module utilized lecture information (i.e., length of the lecture session, video annotations) and the estimates (estimated facial emotions, affective states, and learning affect) to generate comprehensive reports. In this study, report generation modules were embedded for both students and teachers (lecture report). An interactive and live environment was further enabled using a Bidirectional Encoder Representation Transformer (BERT)-based Q&A language model. BERT-based Q&A model takes in the generated narrative report and a text input (pre-defined question from an existing database) either from a student or a teacher and produces a text output (answer) by incorporating basic reasoning. The BERT-based Q&A model helps provide relevant feedback based on the context of the question. Live testing of the framework was carried out using student participants, and survey feedback was used to test the narrative report generation and Q&A module in a laboratory setting. 72.55% of the participants reported that their own experience of the lecture (i.e. comprehension and learning affect experienced) is in line with the model's feedback.

This study found that when students are actively involved in learning, they can exhibit a variety of emotions and affective states that are connected to their learning, and when analysed, can serve as feedback to both teachers and students. This study was based on recent studies in reasoning principles, experiments using machine learning and natural language processing mechanisms, live testing, and questionnaires. Additionally, research demonstrated the link between emotions, affective states, and learning affect as well as how these factors can shed light on a student's learning affect. The examination of affect in e-learning through facial cues estimates, mapping of estimates onto learning affects, generating narrative reports and reasoning using narrative feedback and Q&A for teachers and students is anticipated to facilitate the evaluation of students' learning. We envisage that the proposed framework and future work in this direction will enhance the e-learning platforms and human-machine interaction.

Declaration

I, Christine Bukola Asaju, declare that the contents of this dissertation are my work. This dissertation is submitted for the Doctorate of Philosophy (Computer Science) at the University of the Witwatersrand and has not been submitted to any other university or for any other degree.

Signed: 

Date: May 31, 2023

Acknowledgments

I want to thank my supervisor, Professor Hima B. Vadapalli, for her continuous support, patience, and direction in helping me complete this dissertation effectively. Professor Joel Moitsheki, Head of School, Computer Science and Applied Mathematics, is also thanked for his help and support in finishing this work. My thanks also go to Ms. Marike Kluyts and Dr. Irene Kamara, who volunteered their time to correct the grammatical sections of this dissertation. I am grateful to Ms. Kgomotso Monyepote, Dr. Thama Duba, and Prof. Veronica Ngole-Jeme for their invaluable moral assistance.

I am very thankful to Dr. Samuel Asaju, my husband, for his constant support, sacrifices, and encouragement of my education, as well as to my gorgeous children, Gracious, Wisdom, Testimony, and Joshua. I also acknowledge the inspiration and support I received from my late father, James Olajide Oladimeji, and my mother, Felicia Oladimeji, as well as Mama Mimi Cheyip. The management team of the federal polytechnic Idah is also thanked for their unwavering support of this study.

Publications

Aspects of this dissertation have been published as follows:

- "Estimating Student's Learning Affect: An Approach Based on the Recognition of Facial Emotion Expression". Singh, S., Mujinga, M., Lotriet, H., and Tait, B. (2021). SAICSIT Conference 2021: Proceedings of the South African Institute of Computer Scientists and Information Technologists (pp. 25-36).

- Asaju, C. B., and Vadapalli, H. (2021, November). Affects analysis: a temporal approach to estimate students' learning. In 2021 3rd International Multidisciplinary Information Technology and Engineering Conference (IMITEC) (pp. 1-7). IEEE..

- Asaju, C., and Vadapalli, H. (2022, January). A temporal approach to facial emotion expression recognition. In Artificial Intelligence Research: Second Southern African Conference, SACAIR 2021, Durban, South Africa, December 6–10, 2021, Proceedings (pp. 274-286). Cham: Springer International Publishing.

- Asaju, C., and Vadapalli, H. (2022, November). Adaptive Reasoning: An Affect Related Feedback Approach for Enhanced E-Learning. In Artificial Intelligence Research: Third Southern African Conference, SACAIR 2022, Stellenbosch, South Africa, December 5–9, 2022, Proceedings (pp. 215-230). Cham: Springer Nature Switzerland..

- Spatio-Temporal Reasoning for Enhanced Emotion Analysis: An Investigatory Approach to Strengthen the Current E-learning Platforms. A Poster presented at the SAICSIT 2019, Post Graduate symposium.

- Transfer Learning and its Effectiveness on Facial Emotion Classifier. A short paper presented at SAICSIT 2020 Postgraduate Symposium.

Articles submitted/under review:

- Question Answering Reasoning System for Students' Affect Analysis: A Spatio-Temporal Framework Approach. A journal submitted to the International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKS)(Under review, March 2023)

- Survey paper: Affect Analysis: A Multifaceted Literature Survey of Databases, Methods, and Techniques. A survey paper to be submitted to the Journal of Multimedia Tools and Applications. (June 2023).

List of Abbreviations

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Introduction

The distinction between the real and the virtual world is increasingly diminishing. A greater number of people are using computers to complete various tasks. Online services are used for a variety of tasks, including banking [37], teaching and learning [38], and buying [39]. One industry that has adjusted to the demands and challenges brought on by the inability to conduct face-to-face sessions, which was required because of the global pandemic, is the education sector [40]. Due to these challenges, there are now more possibilities for online teaching and learning [41]. In the online teaching and learning space, the ability to reason about the learning affect of a learner is lacking due to the absence of face-to-face teacher-learner interactions unlike in the traditional classroom settings whereby the teacher can observe and reason about the learning affects experienced by learners through observing various physical cues.

Reasoning is seen to be an intelligent behaviour demonstrated by animate objects. From birth through adulthood, humans have the capacity to consider the connections between what they may visually perceive [42]. They can also draw logical conclusions about the current state of the observed entities' relationship [43]. Psychological studies on reasoning have shown that reasoning involves the formation of actual opinions about objects and their relationship with their environment, based on known facts [43]. Reasoning is the ability to discern meaningful patterns within an information stream, but the inability of humans to discern a meaningful pattern in the stream of data that continuously flows through the senses would result in a world of isolated sights, smells, and sounds, incapable of comprehending or building on experiences across time and space [44]. Reasoning, therefore, makes a meaningful inference from patterns. Being able to recognize facts about objects and actions performed on these objects is an intelligent behavior exhibited by humans [45]. However, this intelligent behaviour by humans has proved difficult for machines in the past decades [1].

Fortunately, recent studies have made a breakthrough in machine learning to determine a meaningful pattern from an information stream. One such study applies static images to a visual-textual relationship, asking questions about objects and their environments that require thinking/reasoning by the machine. For example,

given some static images containing objects of different sizes, colours and shapes, the system is expected to reason and answer questions about the relative position, colour, size, and shapes correctly and this concept was referred to as relational reasoning by [1]. Relational reasoning questions require intelligent thinking about the relationship that exists between objects, while a non-relational question describes the attributes of the objects[1] as shown in Figure 1.1.

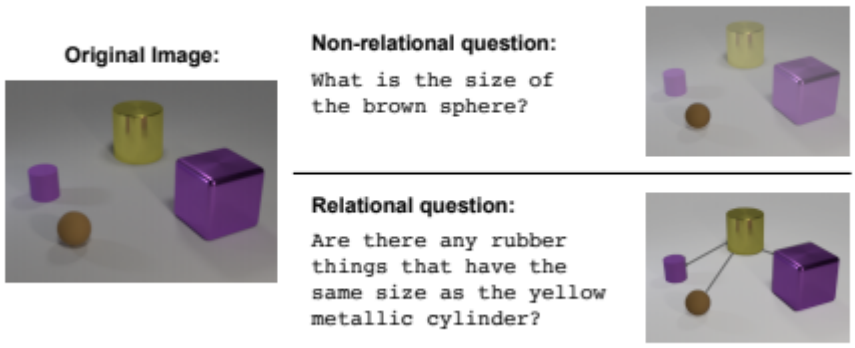


Figure 1.1: An image with relational and non-relational questions[1]

.

Although there is much progress in work related to reasoning using static images and scenes, this area has not been fully exploited. Another emerging perspective on reasoning is Video-based Visual Question Answer (VQA). This VQA reasoning creates video datasets that employ spatio-temporal reasoning applicable to answering reasoning questions [9]. Spatio-temporal reasoning is a type of reasoning in which logical conclusions are made about how objects or entities in videos change as they move through space and time [46].

Over the past few years, the rapid emergence of spatio-temporal reasoning as an important mechanism in deep learning has sparked a great deal of interest in video-related image analytics [47]. Among related works, spatio-temporal-based reasoning is being adopted in many facets of computer vision, such as relative object recognition [9], activity recognition [10], counting and transition recognition [8]. Different architectures such as Temporal Relation Network(TRN) by Zhou et al.[45], Gated spatio-temporal Graph (GSTG) by Tsai et al. [48] and Dual Layer LSTM by Jang et al.[8] have been used for this purpose. The use of these architectures has achieved remarkable state-of-the-art results in specific domain applications that mimic human reasoning.

Consequently, frameworks were developed that show the possibility of a model accepting video input and answering reasoning questions based on these video inputs. Therefore, the existing Visual Question Answering (VQA) was extended to the video domain, giving a new direction to reasoning research. For example, a large dataset for video VQA was developed using a dual approach based on LSTM with a focus on spatial and temporal attention [8]. The research on video question answering was expanded by the investigation of multistep reasoning by the authors in [9]. In their work, a large dataset called Synthetic Video Question Answering (SVQA)

dataset was developed. The dataset developed contained long and structured questions with spatial and temporal relationships between objects. Their system sought to answer relational questions given a video clip relating to the questions. Furthermore, a new attention module to handle spatial attention and a refined Gated Recurrent Unit(GRU) called Temporal Attention Gated Recurrent Unit (ta-GRU) for the attention module were designed. This is because traditional GRU is not capable of capturing long-term temporal dependency.

A new VQA paradigm known as Multi-Question Learning (MQL) was proposed by the authors in [10].The multi-task learning technique, which also gains information from numerous questions and answers for a target video sequence, served as the basis for the VQA. Then, to apply the learned representations of video-question pairs to new questions, they are made more general. An effective VQA framework was also proposed, as well as a training procedure for MQL that utilizes a specially designed attention network that models the relationship between an input video and corresponding questions, allowing multiple video-question pairs to be co-trained.

Amidst various observations and interest in reasoning, it will be of great benefit if the concept can be imbibed in the educational domain, especially e-learning. E-learning is described as the type of learning that is acquired over the internet with the use of computers [49]. However, it was difficult to effectively evaluate the level of learning comprehension on an e-learning platform. In a conventional classroom, teachers estimate the learning outcome of learners by observing facial emotions, eye gauges, body language, and other bodily features [50]. In order to gain feedback from their students on their understanding, teachers also used Q&A sessions [51]. Using physical cues and Q&A feedback, teachers reason and find (relational) reasoning for learners both estimating learning affect and validating those estimates. This helps teachers to obtain an intuitive understanding of the level of participation in the class as a whole. The response is then formulated accordingly using various methods, such as breaks or adjustments in style, pace, and difficult concepts in a lecture, to address the issue [52]. However, this form of reasoning using Q&A sessions to get feedback on learners' understanding has not been explored in e-learning environments.

With the current work in e-learning, emotion is an aspect of the human attribute that plays a significant role in learner learning [53]. With machine learning tools, computers could automatically analyse the learning behaviour of each learner in the e-learning domain. Scotti et al.[53] developed a tool to recognize emotions during the e-learning process, in which they combined qualitative indices extracted from four physical signals, namely: skin conductance, blood volume impulse, electrocardiogram, and electroencephalogram. Their tool was able to detect emotions of stress, relaxation, and engagement during learning [53].

In the domain of e-learning, this study identified that one significant aspect that can imbibe reasoning is in the evaluation of a learner's learning affect. The learning affect of learners was classified through a video using facial expressions by Pan et al. in [12]. Five significant emotions could be identified by their learning expression recognition system. These emotions identified include joyful, surprised, confused, focused, and distracted; and used to auto-evaluate the learners' learning affect. Their work showed that emotions are significant in identifying how learners learn. Facial expressions reflect not only emotions but also mental activities, social

interactions, and physiological signals [14]. However, it was also shown that physical observations in the classroom are not sufficient to correctly evaluate the learner's learning outcomes. Therefore, a learner emotion recognition system to detect the emotions of a learner based on eye and head movements was devised in [14].

In the field of emotional state recognition, the evaluation of learners' engagement in the online environment using facial images was looked at in [54]. Their study used the InceptionNet architecture to measure the level of learner engagement during online learning and reported accuracy rates of 36.5% for boredom, 47.1% for engagement, 70.3 % for confusion, and 78.3 % for frustration. Deep learning methods using FaceNet embeddings and facial landmark points for academic emotion recognition were tested using the DAiSEE dataset, which contains labels for the affective states of engagement, boredom, impatience, and confusion in [55]. Their model was able to identify instances of boredom and frustration that learners felt while taking an online course, with accuracy rates for each affective state of 52.1% and 70.67%, respectively.

Considering the critical role facial expressions and affective states play on education and learner comprehension, it would be beneficial to investigate whether emotions and affective states are explicitly recognized and used for reasoning of a learner's comprehension level. Facial expression and affective states datasets, along with relevant questions will be of enormous importance for the achievement of this goal. Emotion and affective states are not static when it comes to learning [56]. There will be some instances where the learners agree with the content, and there will also be times when they are confused, distracted, or lose interest. Therefore, to effectively analyze the learner learning affects, the video samples containing expressions and affective states will be useful because they contain visual information in both spatial and temporal dimensions.

This dissertation focused on the aspect of generating narrative reports and answering questions based on the exhibited facial expressions, affective states, and learning affect analysis using spatio-temporal reasoning. Spatio-temporal reasoning approach will allow estimations of a learner's learning affect. The analysis of various emotions and affective states plays a very important role in estimating learning affect. Each of the facial cues carries some information that can be inferred as the real state of the individual's mind. This proposed system looked at the ability of machine learning and reasoning principles to estimate a learner's understanding and the possibility of translating the learning affect of such a learner, through reasoning questions and answers. So far, work done in e-learning has not adequately explored an end-to-end pipeline of using estimates (eg; emotions, affective states, learning affects) to devise narrative feedback (reports) and generate answers to queries from learners/teachers by embedding reasoning based on the estimates and also lecture information. Therefore, in this study, we tried to develop a system that covers this aspect of reasoning in e-learning.

So far, in the e-learning domain, researchers have focused on using machine learning tools to classify facial emotions and affective states exhibited by learners and predict their learning affect [54, 55]. The fact that facial signals convey more significant information about the learner's learning affect than other input aspects is the main driver behind the use of emotion and affective states to predict learner

learning affects [14]. However, these studies did not incorporate reasoning through Q&A. This study intends to extend the current setup in the e-learning domain by incorporating the use of facial emotion recognition, affective state recognition, learning affect estimation, feedback report generation for both teachers and learners and reasoning through Question-Answers. Most existing systems of reasoning are based on the use of static images and frames of videos of events that do not consider the temporal understanding of facial expressions and affective states of people. Involving video facial expressions and the affective states of learners in real-time would be a significant contribution of artificial intelligence to the education domain of e-learning. In the traditional classroom learning system, one of the assessment procedures that helps improve the quality of teaching and learning is learning affect evaluation [14, 52, 57]. Teachers evaluate learners' learning attitudes by observing their behaviours during learning and indirectly evaluating their learning affects [12]. This study observed that this form of evaluation is lacking in the e-learning domain.

1.2 Problem statement

Existing e-learning platforms embedded with modules to estimate learner understanding are still in their infancy stage and do not incorporate feedback based on learner cues such as depicted facial expressions and affective states. Estimating a learner's level of comprehension by a combination of his or her estimated facial expressions and affective states, supplemented with narrative and question-answer feedback has the potential in improving existing e-learning systems. Specifically, the following problems were identified:

1. Estimating the learner's comprehension such as positive, negative and neutral learning is a significant challenge in online learning.
2. To use a reasoning approach based on questions and answers phrases as a way of estimating a learner's level of comprehension, using his or her facial expressions and affective states is under-explored and has been gaining traction in e-learning.
3. Existing e-learning platforms are limited by their general feedback and are limited in their ability to incorporate feedback based on depicted facial expressions and affective states of learners using narratives and Q&A feedback.

1.3 Hypothesis Testing

In view of the identified problems, this study formulates a null and an alternative hypothesis as follows:

1.3.1 Null hypothesis

The application of deep learning architectures to process learners' facial behavior and generate feedback narratives and answers using reasoning principles for both

learners and teachers does not provide improvement to existing e-learning platforms in assisting learners and teachers.

1.3.2 Alternative hypothesis

Applying deep learning architectures for processing learners' facial behavior and generating feedback narratives and answers using reasoning principles for both learners and teachers will provide an improvement to existing e-learning platforms in assisting learners and teachers.

1.4 Research Questions

The questions to be answered in this study are given as follows:

1. To what extent can facial expressions and affective states help determine learners' comprehension in e-learning by estimating their learning affect?
2. To what extent can a cascade of spatio-temporal models and their estimates generate useful narrative feedback for both learners and teachers?
3. To what extent can a language model reason and answer questions based on narrative feedback reports generated from estimates of facial expressions, affective state, learning affect and lecture annotations?

1.5 Research aims

To respond to the research questions, this study aims to:

1. Adopt spatio-temporal model (CNN-BiLSTMs) to investigate its capacity to handle facial video input.
2. Perform learning to affect analysis using estimated facial emotions and affective states.
3. Generate a narrative feedback report for teachers and learners based on estimates of facial emotions, affective states, and learning affects.
4. Investigate and explore the use of transfer learning by fine-tuning the pre-trained BERT model to generate answer phrases in an educational context.
5. Generate feedback answers using fine-tuned BERT model for question texts submitted by teachers or learners, using narrative feedback reports and lecture video annotations.
6. Investigate the performance of the proposed pipeline using live testing while adhering to ethical considerations.

1.6 Objectives

The objectives are to:

1. train, validate and test a CNN-BiLSTM model that can recognize facial expressions of happy, disgust, anger, fear, neutral, surprise, and sadness, using samples from DISFA+ dataset.
2. train, validate and test a CNN-BiLSTM model that can recognize affective states of boredom, confusion, engagement, and frustration, using samples from DAiSEE dataset.
3. map the estimated emotions and affective states to relevant learning affects such as positive, negative and neutral categories based on literature survey, and verify these mapping using a cross-model approach.
4. generate a narrative feedback report using a template-based approach and incorporating estimates of emotions, affective states, and learning affect including available lecture video annotations.
5. generate answer phrases using the BERT model to questions posed by both teachers and learners.
6. test a cascade of the trained models (emotion estimator and affective state estimator) integrated with the narrative feedback and question-and-answer model on live sample data obtained from the University of the Witwatersrand students.

1.7 Expected Impact of the Study

An improved learning affect detection has a significant impact on learning [58]. This study is expected to enhance e-learning platforms by automatically estimating a learner's learning affect using facial emotion and affective state recognition, and mapping estimated emotions and affective states onto learning affect using CNN-BiLSTM cascade for handling spatial and temporal video streams. Furthermore, generating narrative feedback based on the estimated emotions and affective states, learning affect mapping of the estimated emotions and affective states, and video annotations. In addition, incorporating a BERT-Q&A language model by using the generated narrative report along with reasoning questions based on estimated facial cues with corresponding answers to reason about the learning outcomes, using this to further improve learners' understanding. The outcomes of this study are expected to shed light on the successful use of existing state-of-the-art machine learning and natural language processing architectures for improving e-learning platforms. The study is expected to have some specific impact on, but is not limited to, the following aspects of e-learning:

1. To help teachers who are aware of the affect of learning of learners better deliver lessons [59].
2. To enhance quick and effective identification of learners with learning disabilities [60].

3. It is expected to provide an effective evaluation of teaching-learning processes by teachers and learners.
4. Teachers can improve the effectiveness of their offerings (modules) by considering the affective domain [61].
5. An improved learning affect detection is expected to help teachers and learners to know the stages of their learning at which a learner is struggling in a course and provide solutions that can help prevent eventual dropout of learners. Understanding, predicting, and mitigating dropout behaviour among gifted learners may require the use of learner learning as a construct [62].

1.8 Novelty of the Study

Even though work has been carried out on affect analysis in e-learning, the analysis through the combination of emotions and affective states, mapping of estimated emotions and affective states onto learning affects such as positive, negative and neutral, generating narrative feedback based on the estimated facial cues and video annotations and reasoning using the narrative feedback and Q&A feedback for teachers and learners has not been explored. Therefore, this study identifies the absence of reasoning through the analysis of affect in e-learning using facial cues estimates, mapping of estimates onto learning affects, along with video narratives to generate a narrative report and further use the narrative feedback generated from the estimated facial cues and learning affect mapping with Q&A feedback for teachers and learners as the key gap. The study, therefore, explores the integration of narrative feedback and Q&A module into a cascade model that estimates both facial emotions and affective states of learners, mapping them onto learning affect for reasoning. We expect that this study will serve as a foundation for advancement in the e-learning domain, particularly in the understanding of learners' learning affect and being able to effectively evaluate learners' learning responses and provide appropriate feedback to both teachers and learners themselves.

1.9 Chapter Summary

Research on the estimation of emotions and affective states, mapping estimates onto learning affects, and their use in generating feedback as both narrative reports and Q&A were explored. The key gap that led to the work was identified, which included the lack of a complete framework that includes narrative feedback based on learner facial cue estimates, learning affect estimates, generative narrative reports for teachers and learners, and reasoning through questions and answers in an e-learning environment. Problems that needed to be addressed, the research questions, as well as the research aims and objectives, were included in this chapter. The expected impact of the study was also discussed. As feedback on a learner's comprehension is critical to a teacher during a teaching and learning session, a learner also desires to understand his or her learning affect during learning; this work provides a springboard for accomplishing this.

The remainder of this dissertation is structured as follows: Chapter 2 presents a review of related work. Existing literature on reasoning using static images, spatio-temporal reasoning with video images, traditional classrooms, and e-learning is presented. Current techniques for analyzing affects and feedback mechanisms were also discussed. Chapter 3 sheds light on background mechanisms and techniques, including an overview of facial expressions, affective states, e-learning, and datasets used for the proposed study. The state-of-the-art feature extraction, classification, and natural language processing mechanisms, along with evaluation metrics, were discussed. Chapter 4 presents the proposed research, including a detailed description of each of the components in the pipeline. Chapter 5 sheds light on the results reported from the different experiments carried out in this study. Final mapping of emotion and affective state onto learning affect, narrative report generation, Q&A generation, and live test results are discussed. Chapter 6 concludes this dissertation.

Chapter 2

Review of Related Work

2.1 Introduction

From chapter 1, it can be inferred that studies around reasoning has attracted a great deal of attention in the past years. Part of the work in this area of study is reasoning about static images and dynamic images, such as videos. The latter is known as spatio-temporal reasoning. Spatio-temporal reasoning is reasoning that deals with data with the characteristics of both space and time. This chapter discusses the literature on spatio-temporal reasoning and its application to the e-learning domain.

The literature is divided into four main areas. These include reasoning using static images, reasoning using video streams, and work carried out in the education domain followed by work that has been carried out on the use of question-answer systems to demonstrate reasoning. In this review chapter, existing work along with areas of application have been thoroughly covered to establish the gap that necessitated this study.

Section 2.2 examines the studies on reasoning, followed by 2.3 which establishes the significance and use of spatio-temporal reasoning. Section 2.4 sheds some light on the behaviour studies carried out around traditional classroom and e-learning domains. Section 2.5 discusses the different techniques used for affect analysis, while we discuss the feedback mechanism in 2.6. Section 2.7 then provides a motivation for the current work, followed by 2.8 which provides a summary of the chapter.

2.2 Reasoning

Reasoning is an important, yet difficult task to understand interactions in visual scenes [2, 3]. For example, "boy, climb, bicycle" is faced with a complex relation of "climb" between "boy, bicycle". Many studies have explored the domain of still images, which involves the use of pictures of scenes or objects as input and reasoning questions and answers are attached to give a reasoning output for the images. Examples of such questions are *"what colour is the other object that is the same shape as the large brown thing?"* [4], *"what size is the cylinder that is left of the brown metal thing that is left of the big sphere?"* [1]. While visual relationship detection such as:*"woman rides a bicycle", "man kicks the ball," "sky is above the water"* [3]

and many more reasoning questions were intelligently given reasoning answers by the machine.

Some work in reasoning includes the use of a Relational Network (RN) that is capable of solving reasoning puzzles in the field of artificial intelligence, to solve reasoning problems as proposed by Raposo et al. in [2]. This RN module implementation is an architecture geared toward solving object-relation reasoning. RN was used to classify a static scene and then combined with Memory Augmented Neural Network (MANN) to show that neural networks could be supported by other architectures to reason efficiently given a new scene. The RN-MANN model was further experimented with a chaotic scene by augmenting RN with an architecture that will transform an unstructured scene directly into a representation that has a scene description matrix. This was achieved by creating a dataset of scene descriptions and their related images. The results were reported by using targets that explicitly describe the relations between objects and accuracy of 97% was reported [2]. Figure 2.1 shows the relational network with a memory-augmented neural network.

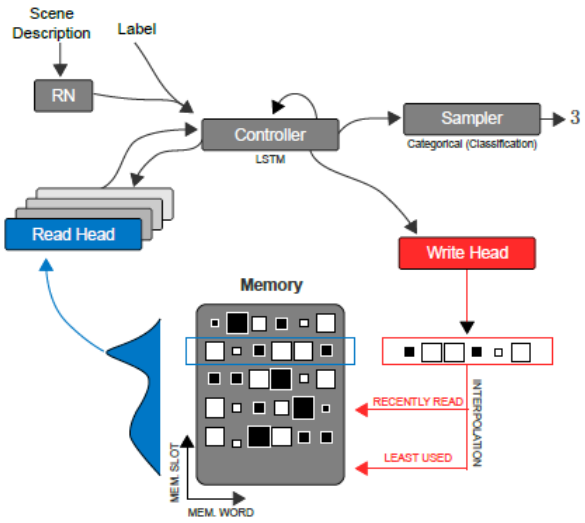


Figure 2.1: RN with Memory Augmented Neural Network[2]

Even though powerful Convolutional Neural Networks (CNN) do not have a general ability to tackle reasoning questions, they can do so when combined with RN [1]. The RN described in the work of [2], operates on objects, but does not thoroughly operate on images or natural language. Therefore, the authors of [1] applied a CNN to read the pixel of a set of objects. CNN is a deep learning algorithm that can take an image as input, then assign weights and biases to the various parts of the object in the image. CNN can distinguish each component of the image to produce a set of feature for the RN [63]. Questions can be processed using LSTM, a Recurrent Neural Network (RNN), which has been used in the deep learning domain and is still widely used for solving classification, clustering, and data prediction problems, especially time series and textual data [64].

Furthermore, a Compositional Language and Elementary Visual Reasoning Dataset (CLEVR) was used to evaluate a variety of modern visual reasoning systems, providing different insights into what they can and what they cannot do in [65]. The model

that experimented with three reasoning tasks includes VQA, text-based Q&A, and dynamic physical systems. The RN takes into account relationships between every pair of objects, subject to the embedding of a question, and integrates all of these relationships to provide an answer to the question. RN module inclusion in relatively simple CNN and LSTM-based VQA architectures increased performance of the CLEVR data set from 68.5% to 95.5% and reported state-of-the-art superhuman performance [1]. The VQA architecture is shown in Figure 2.2.

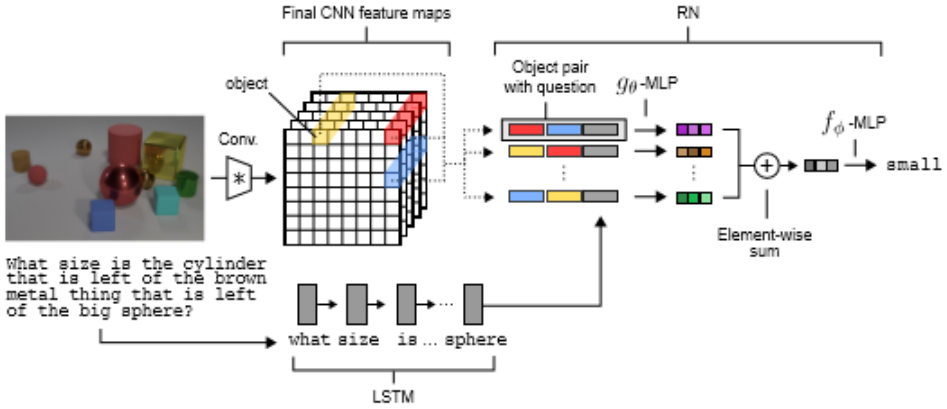


Figure 2.2: The Visual Question Answer Architecture [1]

Another related work in this area of reasoning is that of Dai et al. [3]. Their research described a deep relational network model that takes into account the statistical dependencies of objects and their relationships. This method was able to recognize many types of relationships such as relative positions (e.g. behind, above) actions (e.g. eat, sit), functional (e.g. part of), and comparison (taller than, etc.). The primary task was to identify all visual relationships from a given image and to draw conclusions about the triplet, i.e, subject, relationship, and object (e.g., man, eat, apple) [3]. Their work comprises three stages, namely: object recognition using a faster region CNN, pair filtering that produces the informative pair of objects, and joint recognition that sends the filtered pair of an object into the recognition module, which in turn produces a triplet as output [3]. Three tasks were also examined. First was predicate recognition, which focuses on the recognition of subject and object. Second, union box detection in which all three elements, i.e., the subject, the relationship, and the object, are predicted. Lastly, the two-box detection task, which predicts the subject and the object in the scene, but requires the Intersection Over Union (IoU) was examined. The performance for both the subject and object was above 0.5. When the model was tested in the Visual Relationship Detection (VRD) and Subject Visual Genome (SVG) datasets, an accuracy of 81.90% was reported for VRD and 91.26% for SVG [3]. Figure 2.3 illustrates the visual relationship detection model framework.

Feature-Wise Linear Modulation (FiLM) approach was introduced in Perez et al. [4]. FiLM layers were shown to achieve good results in answering image-related questions in their work. The FiLM model consists of a FiLM generating linguistic pipeline and a filmed visual pipeline using Gated Recurrent Unit (GRU), a type of LSTM without an output gate, and a CNN to process questions and images. The

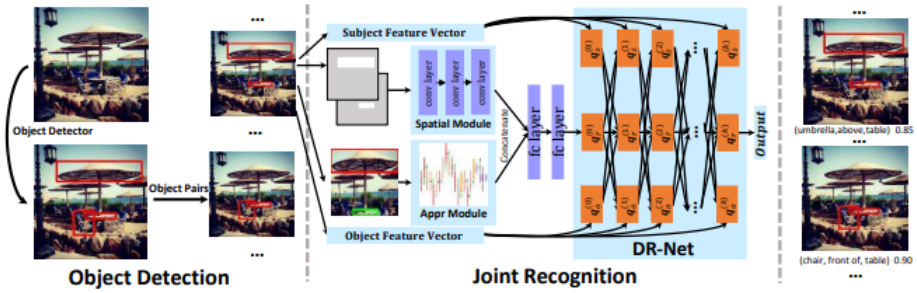


Figure 2.3: A framework for visual relationship detection [3]

.

model was tested on visual reasoning with the CoCLEVR task and what the film learns was analyzed using trained film models. The model was also tested with the CLEVR human task to establish how well the model responds to challenging and general questions. Both outputs applied on a FiLM network gave a state-of-the-art result. The CLEVR accuracy reported was 96.9%, while that of the CLEVR-Humans data was 75.9%. By and large, FiLM performed well in learning general concepts [4]. Figure 2.4 gives the FiLM generator model framework.

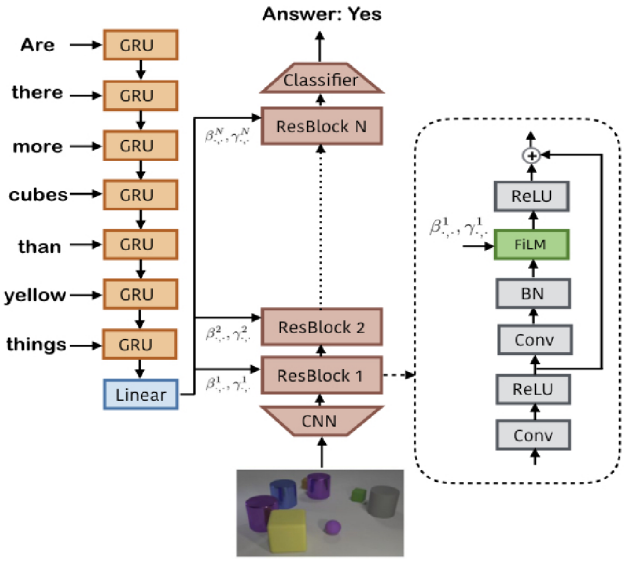


Figure 2.4: The FiLM generator model [4]

Simple plug-in relational module proposed by [1] was too expensive, therefore with this notion, the authors in [66] proposed DenseNet, a method that tries to avoid the plug-in module simply by adding a dilated convolution to a robust architecture of CNN. The experiment was carried out using the SORT-of-CLEVR dataset to test the accuracy of relational and non-relational abilities of different models, of which the Dilated Dense was also a part. This was achieved by first reproducing the work of [1] and denoting it by CNN+ RN. Second, the same network was trained, but this time the RN module was replaced by a 2-layer Multi-Layer Perceptron (MLP), which is

a class of feed-forward artificial neural network, which was denoted by CNN+ MLP. It was discovered that the network with a relation module CNN+RN reported a higher accuracy than the one that does not, that is, CNN+ MLP. CNN+ RN gave an accuracy of 87.7%, while CNN+ MLP gave an accuracy of 65.4%, indicating that the relation module is an essential component of any reasoning network. In the second part of their experiment, a dilated DenseConvolutional Neural Network (DenseNet) and an ordinary DenseNet was used. The Dilated DenseNet was augmented with a 2-layer MLP, (Dilated DenseNet + MLP) and reported an accuracy of 83.7%, and a DenseNet without dilation(DenseNet+ MLP) to observe the importance of dilation in reasoning. It was observed that this network did not perform as the network with dilation, and an accuracy of 66.7% was reported [66].

In another related work, a model for inferring relationships whose task was to reason about the relationship between an object and a subject, distinguishing between the actions performed by the specified subject on the object and on other subjects in the same scene was developed by Krishna et al. in [5]. CNN was utilized in their model to extract the image feature, which was then used to construct an initial feature vector for the subject and the object independently. These feature vectors were utilized to distract attention away from the subject and toward where the item is meant to be. By concentrating on the redirected area, the visual aspects of the new estimation of the item were modified. At the same time, an inverse shift from the starting point to the subject was found. The two predicate shift modules communicated with one another to localize the subject and object. The performance of the model was evaluated using three datasets made up of visual relationships, namely: CLEVR, Visual Relationship Detection (VRD) and Visual Genome (VG). Each dataset presented distinctive features of 33%, 60.3% and 61% of relationships respectively, and distinguished entities of the same type with many occurrences. Second, the model was tested in the absence of one of the input reference relationships (subject, predicate, and object), and it was shown that the model could still differentiate between entities and could locate entities from the new scene [5]. Figure 2.5 shows the framework presented in [5].

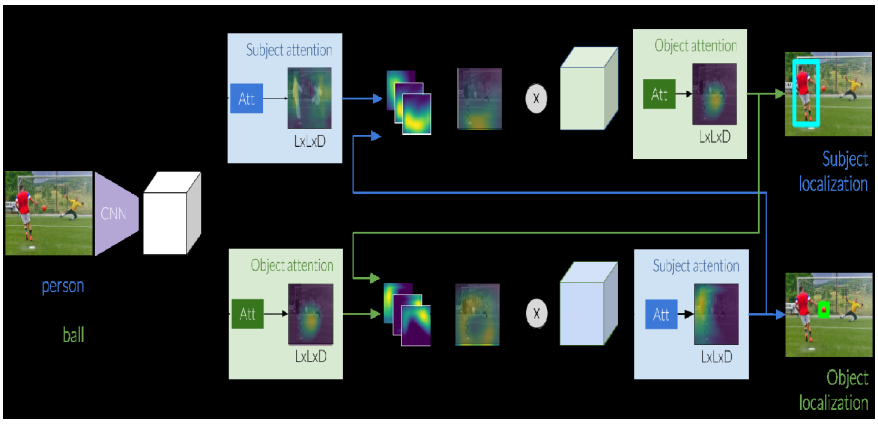


Figure 2.5: The Referring relationships model [5]

RN is a general-purpose module that operates on a graph representation of objects [67]. Graphs are made up of vertices or nodes connected by edges. The module

can assist any neural network model with the ability to perform many-step reasoning. A model to solve tasks that involve multi-step reasoning such as answering complex questions about the relationship between many objects or solving puzzles that require many steps of reasoning was developed by the authors in [67]. "Babi and sort-of-CLEVR" datasets which require a few steps, "Pretty CLEVR" which requires up to eight steps, and Sudoku which requires more than ten steps were adopted from [1]. These methods were adopted to solve the Sudoku puzzle and an accuracy of 96.6 % was reported. Another work in reasoning that focused on the underlying reasoning process and included a compositional model for VQA, which comprised a program generator—a software program that allows individuals to create programs of their own with little or no effort and programming knowledge, constructed the reasoning process to be performed using a standard LSTM. The compositional model for VQA also consists of an execution engine that provides the answer by implementing neural networks, which is the neural module network. The model receives an image as input along with a visual question regarding the image, and selects its answer from a predetermined range of options. The model was evaluated using the CLEVR dataset, and it was established that standard VQA methods do not produce a good result on this dataset. Therefore, it was proven that the combination of program generators and execution engines works much better on CLEVR than other methods. An accuracy of 96.7% was reported on their program generator and execution engine model using the CLEVR datasets [6]. Figure 2.6 shows an overview of the program generator and the execution engine.

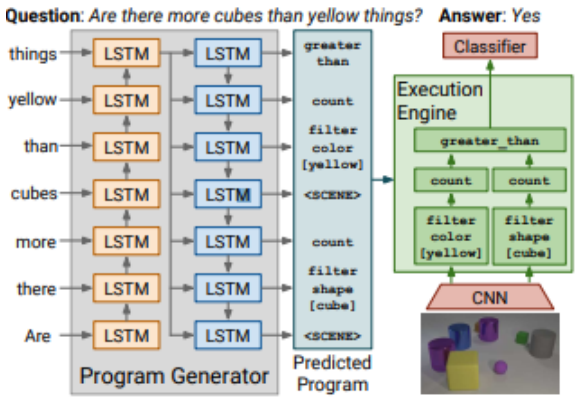


Figure 2.6: The program generator and the execution engine [6]

A Neural Module Network (NMN) architecture for VQA, which consists of jointly trained neural modules, using End-to-End Module Networks (N2NMNs) was proposed by Hu et al. in [68]. The NMN proposed learns to reason without a parser (grammatical analysis of the text input). NMN directly predicts the instance-specific network layout, which means that it takes textual inputs and applies them directly to an image. The model was first analyzed using relatively small SHAPES datasets as proposed by the authors in [69]. This dataset contains questions about simple arrangements of shapes. An accuracy of 100% was reported. Subsequently, the model was tested on two large-scale datasets, namely CLEVR and VQA. Analysing the results on CLEVR dataset, the authors deduced that their neural module-based

methods were able to reason about complex questions in large datasets and reported an overall accuracy of 83.7%. On VQA, when tested using real images, compared to other methods that used modular structures, the authors reported an accuracy of 64.2% [68].

The works presented so far on reasoning have utilised static images. The utilization of video input for reasoning has been identified as an extension of using static images. The output result of a reasoning task that involves answering questions on the various video inputs is referred to as spatio-temporal reasoning. Reasoning questions in spatial and temporal scenes have been explored so far in the field of machine learning. For example, in a dynamic scene, questions such as; *"how many times is an action carried out?", "what does the bear do right after sitting?"* [10], were asked through video input, and reasoning answers were given by their system. Moreover, different deep-learning approaches have been explored to achieve remarkable results, and several authors have contributed to this task. The following section reviews work on spatio-temporal reasoning.

2.3 Spatio-Temporal Reasoning

Spatio-temporal reasoning is an area of artificial intelligence that draws from other related fields such as cognitive science and cognitive psychology. Spatio-temporal reasoning involves a higher order of reasoning, which is used for operating and comprehension of time and space [70]. Generally, spatio-temporal reasoning is the ability to visualize the relationship between three-dimensional objects and to detect their transformations over time. Spatio-temporal reasoning is therefore regarded as the understanding ability that plays an essential role in many fields of learning such as architecture, engineering, computer science, and mathematics. Invariably, spatio-temporal reasoning is reasoning about the movement of objects through space and time [70]. Chen et al.[71], also reiterated that spatial and temporal relations are incorporated to represent expected dynamics in a scene. Wolter et al. [7] proposed a theory stating that "space", which is a spatial dimension, is always constant, although regions may change from their present position at different times. Therefore, spatio-temporal relations is described as a Cartesian product of space and time structures [7]. Figure 2.7 illustrates the theory as proposed by Wolter et al.[7].

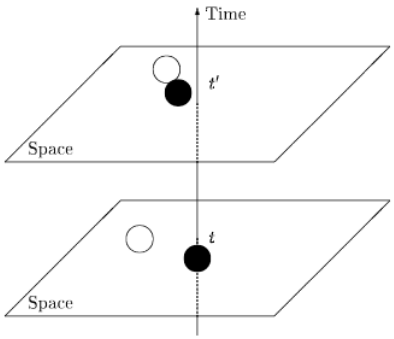


Figure 2.7: Spatial region moving with time [7]

Due to the spatial-temporal nature of video data, understanding visual relationships in video scenes has received limited attention so far [10, 8]. The limited attention was attributed to the fact that spatio-temporal reasoning involves modelling a more comprehensive set of visual relationships [48]. Consequently, this led to further studies in the area of spatio-temporal reasoning.

Video data is made up of visual information that is spatial and temporal. Three tasks on video-based reasoning were proposed by Jang et al.[8]. One of these tasks uses spatio-temporal reasoning from videos to appropriately answer questions. Second, a new large-scale dataset for video VQA called Tumbir Graphics Interchange Format-Question & Answer (TGIF-QA) was introduced, which is an extension of the existing visual question-answering. A dual approach based on LSTM with special and temporal attention was also used, and their work was compared with other existing VQA techniques. Specifically, the video (VQA) was to determine the state transition, detect a recurring action given its count, and count the repetitions in a given action. This indicates what has happened before and after a specific action using a short video caption Graphic Interchange Format (GIF). The method of solving this task that requires spatial-temporal reasoning from videos was described. Question-Answer pairs from the video captions were also generated. They reported a different result in terms of visual question-answering tasks with accuracies of 46.44% for action reasoning, 63.79% for state transition and 47.79% for frame reasoning, respectively. Figure 2.8 shows the framework for the ST-VQA model.

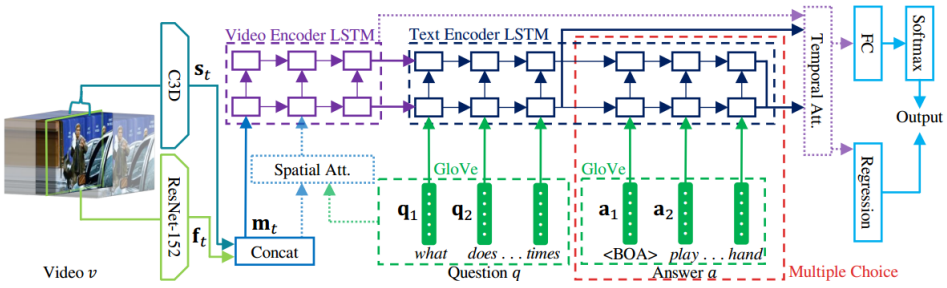


Figure 2.8: The ST-VQA model for spatio-temporal reasoning [8]

Work in the field of spatio-temporal reasoning was extended due to the fact that the existing video VQA datasets are not suitable for multi-step reasoning [9]. These videos lacked logical structures and were language biased and as a result of this, a large dataset known as SVQA was created by the authors of [9]. SVQA dataset consists of long, structured questions with spatial and temporal relationships between objects. Further more, a new video model called VideoQA (VQA) was created by the authors of [9]. Synthetic video was chosen in their work because of its particular spatial and temporal relations among objects, which the authors felt was suitable for multi-step reasoning. Subsequently, a new attention module was designed to handle spatial attention, and a refined GRU called a ta-GRU. This is because traditional GRU cannot capture long-term temporal dependency. The result of their experiment showed the effectiveness of this method compared to existing models and reported 53.53% accuracy on QA frame. Their model framework is shown in Figure 2.9

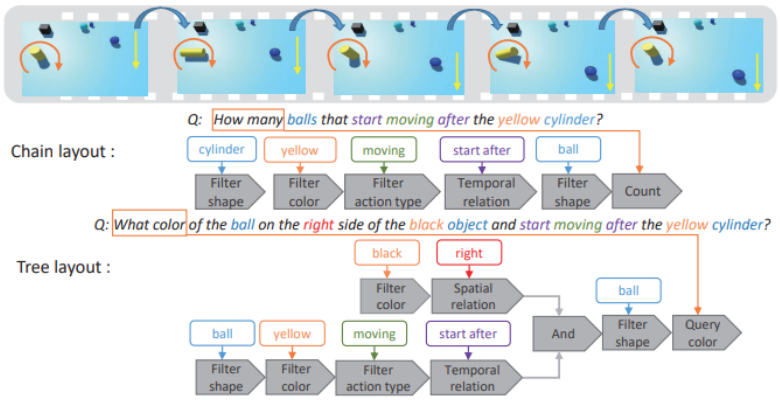


Figure 2.9: The Synthetic Video Question Answer model [9]

Additionally, a multi-stream end-to-end trainable neural network for multimodal video Q&A was introduced by Jei et al. [10]. One of their contributions to this field of spatio-temporal reasoning is the development of an extensive dataset of videos with a very long length of playing time. The features of the video frames were extracted and Faster Region-Based Convolutional Neural Network (R-CNN) was run on the visual genome to detect the objects in the video frame and the attributes in each of the frames. The use of the detected object label as input to an image captioning system was found to provide performance comparable to that of directly using the CNN features. Therefore, these authors used detected labels as input and BiLSTM was used to encode textual, while visual sequences and Global Vectors for Word Representations (GloVe) were used for word embedding. Test results reported an accuracy of 68.48% using all available sources of subtitles, visuals, and questions [10]. Figure 2.10 illustrates the multi-stream model for multi-modal video QA.

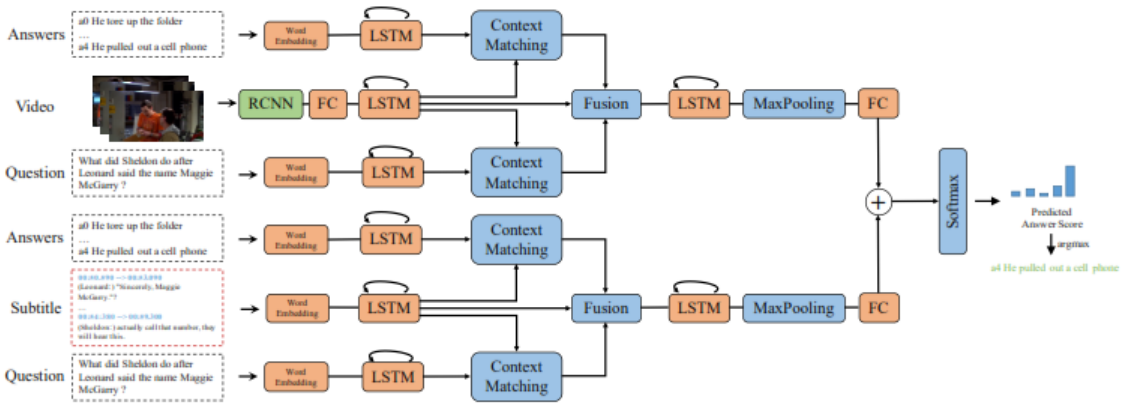


Figure 2.10: The multi-stream model for Multi-modal Video QA [10]

The ability to link significant changes in objects or entities with respect to time is known as temporal relational reasoning and is regarded as an intelligent behavior [45]. As a result, a network module known as Temporal Relation Network (TRN), which learns to justify the time dependency between video frames at various time scales, was proposed by Zhou et al. in [45]. The authors focused on video activity recognition to describe the steps of an event. Although they did not attach a Q&A module to their work, it was inspired by the reasoning module for VQA and reported 94.78% accuracy using multi-scale TRN.

A fully connected spatial-temporal graph called the Gated Spatio-temporal Energy Graph was proposed by Tsai et al. [48]. The Gated Spatio-temporal Energy Graph was used to model the spatial and temporal structure of the relationship in videos. Each node in this graph represents a distinct entity, and the edges show their statistical relationships to one another. The system uses reasoning on a fully linked spatio-temporal network with nodes S, P, and O representing the subject, the predicate, and the object, respectively, to predict the output connection given an input image of videos. They produced cutting-edge results using this methodology for the three reasoning tasks of detection, tagging, and recognition. It has been suggested that videos can assist in resolving issues with temporal correlation and ambiguity in a set of relationships. On these three relationship tasks, a cutting-edge performance was noted [48].

An Object Relation Network (ORN) that was described as a neural network module which can reason between instances of objects detected through space and time was proposed by Fabien et al. [72]. The ORN could reason between the pairwise relationship between objects detected at a different moment in a given time thereby identifying the movement of objects and their interactions. Objects and features were detected and computed by the object head. Based on a study by Pickup et al.[73], who reported that activities are completed over a period of time and that their model had an accuracy of 40.89%, we deduce that the logic behind the activity in the video is temporal.

A method for Human Activity Recognition (HAR) that collects information at different resolutions using CNN with changing kernel dimensions and BiLSTM was proposed by Nafea et al.[11]. The innovative aspect of their study is the effective selection of the appropriate video representation as well as the effective extraction of spatial and temporal information from sensor data using conventional CNN and BiLSTM. Their proposed methodology makes use of wireless sensor data mining (WISDM) and UCI datasets. The results show that the proposed framework improves HAR effectively. It was discovered that, in contrast to other known methods, their proposed method improved accuracy, achieving a better score in the WISDM dataset than the UCI dataset. The CNN-BiLSTM model reported an accuracy of 98.53% in WISDM and 97.05% in UCIHAR. In terms of spatial and temporal features, this was an acceptable level of accuracy. Their framework is shown in Figure 2.11.

Research carried out in spatio-temporal reasoning adopts the use of video datasets instead of static image datasets for reasoning tasks. Video datasets are preferred because videos possess the properties of spatial and temporal behaviour. Video data comprises a series of still images that give the appearance of motion when viewed in

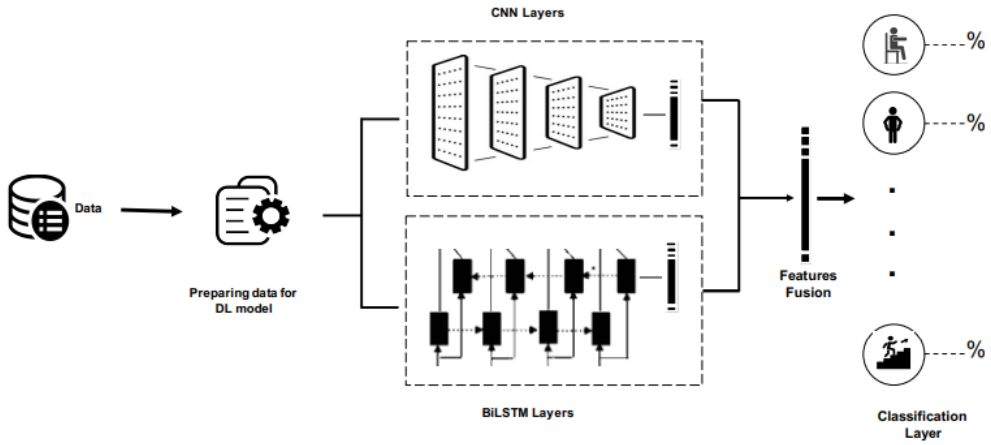


Figure 2.11: HAR model using CNN-BiLSTM Framework [11]

a sequence of speed, giving the appearance of motion [74]. Reasoning in real-time therefore requires a video input domain, dependent on space and time.

2.4 Behaviour Studies: Traditional Classroom and E-Learning Domain

The education domain has incorporated artificial intelligence into its systems in several ways. These include, among others, educational resources such as e-Libraries, intelligent computer-assisted instruction, various databases for learners and instructional roles. The evolution of computers has significantly impacted on learning platforms. Due to the continuous increase in virtual education systems worldwide, computer-mediated education is indispensable for the future [75]. Jones et al. in [76], also stated that the instructional role is the application that has received the most attention of all the various uses of computers in education.

Even though e-learning as an aspect of computer application in education domain plays an important instructional role in education system, understanding the comprehension of learners has not been much explored. E-learning is a computer-based instructional tool that allows the learner to learn anywhere, anytime, on the Internet [76]. Considering this aspect of education and exploring it to reason will be of enormous benefit to both learners and teachers.

Learning is an emotionally charged event [29]. The learning process is progressively augmented by computers and the internet [53]. Therefore, integration of emotion into the computer-augmented system can result in improved learning behaviour. Academic emotions are defined as achievement feelings such as frustration and anxiety, social emotions such as pride and jealousy, topic emotions such as empathy for a protagonist, and epistemic emotions such as confusion and surprise [77, 78, 79]. Confusion, frustration, boredom, engagement, curiosity, anxiety, delight, and surprise are some of the feelings learners communicate through their facial expressions [80]. Generally, the attempt to accurately access learners' impact

has led to diverse machine learning and computer vision studies.

Assessment was described as the "unseen villain" in many educational discussions [81]. Invariably, the education obtained is defined by the assessment. The authors in [82, 83] pointed out that the assessment of learner learning outcomes has become a key issue for various institutions of learning in terms of accreditation, accountability, and performance indicators. Therefore, it is necessary for a teacher to be able to quickly access the conditions of the learner so as to teach effectively and to address the demands of learners during a class [52]. Teachers use a variety of strategies to assess learner understanding in the classroom, including facial signals, body movements, gaze, and other methods [50]. Teachers continue to strive to better understand the affect of learners during learning.

2.4.1 Learning Affect

Affect is a term used in psychology to describe how people express their emotions and moods [84]. Affect is a broad concept that encompasses a variety of emotions and moods. Affect is often used in the educational area to refer to a wide range of non-cognitive variables, such as emotions, self-concepts, beliefs, and motivations [85]. When learners are learning, they feel a variety of cognitive and affective emotions. Compared to a system that overlooks the affect of learners, a computer system that is responsive to the affect of learners in a learning environment is projected to boost learners' motivation and learning gains [86]. Generally, as part of reasoning, the assessment of a learner, to measure the success of teaching and learning, is one of the most important tools used in the teaching and learning process. Both in the classroom and online, affect is an important aspect of learning activities. learners' affect is defined by the attitudes, interests, and values that they exhibit and absorb in school[87]. Affect was referred to as emotions and moods by Pekrun et al. in [78]. The learning affect of learners in a classroom setting was categorised through their facial emotions and expressions by the authors in [12]. Classroom teaching and learning affect were auto-evaluated and five significant learners' emotions, based on their learning expression recognition system were identified. These included joyful, surprised, confused, focused, and distracted.

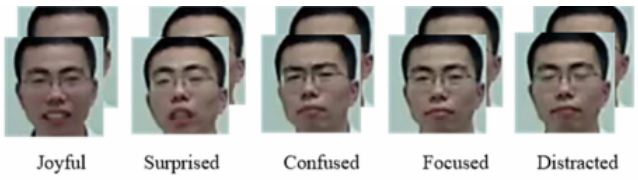


Figure 2.12: Five types of emotions when learning [12]

2.4.2 Importance of Detecting Learning Affect

The attitudes, values, and interests that learners demonstrate and learn in school are referred to as learners' learning affects according to Popham et al. [88] and have been shown to have a significantly higher influence on learners' post-secondary life

than their academic performance. Learning affect detection is important for the following reasons:

1. It allows teachers to make a conclusion about the learner, which aids them in planning their lessons [89].
2. It supports the creation of affect-responsive learning environments, which help students control their emotions and have more enjoyable learning experiences. [90].
3. It helps improve learning outcomes by recognizing and responding to affective states that are detrimental to learning, as well as inducing those that are helpful [91].
4. It helps in enhancing one's understanding of learning processes. This is achieved by providing non-intrusive, moment-to-moment tracking of ephemeral affect-cognitive interactions that are outside the radar of conventional self-reporting measures [91].
5. It acts as a magnifying glass to get a closer look at the different learning phenomena [91].
6. permits teachers to make a conclusion about the learner, which helps them prepare their lessons [89].
7. It benefits learners, as teachers who are aware of their learners' affective states offer better-targeted instruction that leads to successful learning experiences [92].

2.4.3 Affect Analysis: Approaches

The fundamental purpose of a system that can recognize and analyze affect is to use input signals from several modalities to convey information about an interpreted emotion. To carry out affect analysis in the online learning system, several methods have been used. Methods such as basic emotions, affective states, positive emotions, negative emotions, neutral emotions, engaged, not engaged, comprehended, not comprehended, and a variety of different analyses have all been conducted. Furthermore, input signals such as physiological signals, face, voice, text, body language and multi-modal systems have been utilized. Various categorizations, databases, as well as techniques and results, have been presented in the last few years to meet the goal of improving learner learning. These are discussed in 2.4.4 to 2.5.3.

2.4.4 Affect Categorizations/Mappings

Human affects can be classified using two fundamental principles known as discrete basic emotion categorization and dimension technique[93, 94, 95]. Emotions are referred to as discrete because they can be distinguished fundamentally from one another [96]. Several authors have presented their views concerning the discrete categorization/mapping of affect. Some of these include Sauter et al. [97], who presented an outline of a study on the expression of positive emotions. Their research

classified emotions into clusters based on shared criteria. Happy, relief, awe, and interest were identified as epistemologically positive emotions in the study.

In the dimension approach, emotions are classified into three dimensions, which include valence, arousal, and dominance [94]. The term "valence" refers to the degree to which a feeling is negative or positive [98]. Valence is divided into two categories known as low valence or negative valence emotions (fear, tenseness, anger, sadness, boredom, etc.) and high valence or positive valence emotions (joy, happiness, happy, calm, delighted, excitement, etc.) [95, 99]. On the other hand, arousal is described as the intensity of an emotional state, also reflecting the level of enthusiasm or apathy [95, 100]. Arousal is further subdivided into two groups, which are the negative or low-arousal emotions (relaxed, tired, bored, calm, depressed, etc.) and the positive or high-arousal emotions (excited, angry, tense, frustrated, happy, etc.). Action Units (AUs) were being matched to learners' learning experiences in a study by Kapoor et al. [26]. This is because the facial action units define how the face deforms as a result of facial muscle movements [101] and the combinations of different action units define different emotions [102]. Their findings show that a smile (Action Unit (AU)6 + AU 12), tightening eyelids (AU 7), expanding eyes (AU 5) and rising eyebrows (AU 1 + AU 2) show that the learner is engaged and enthusiastic about the learning phase or activity. Their findings further showed that lowering eyebrows (AU 1 + AU 4), wrinkling nose (AU 9), and lower lip corners (AU 15) indicate a lack of motivation.

A mapping mechanism based on classroom observations and instructors' responses to questionnaires where learners' facial movements were mapped to their comprehension levels was proposed in [25]. Wide-open eyes and raised brows denoted successful or positive learning, while closed eyelids, forehead lines, and curled lips denoted unsuccessful or negative learning. A smile or neutral face denoted an unclear learning status. The authors in [25] further categorized anger and sadness as negative affect, while fear, surprise and happy were categorised as positive affect. Students' emotions of anger, happiness, surprise, and neutral were classified as positive affect according to studies in [12], while disgust was ascribed to a negative affect. The changes in learners' facial emotions were estimated and related to learning affect in [27]. Negative learning affect was mapped to anger, disgust, and sadness, whereas positive learning affect was linked to fear, happiness and surprise. Neutral emotion was mapped to a neutral learning affect. These mappings and categorisations are expected to assist researchers in better analyzing the affect estimation results. Prior to the various mappings, several datasets have also been applied by different authors for the analysis of affect. These are described in the next section 2.4.5.

2.4.5 Existing datasets available for Affect Analysis

Various datasets have been used for the study of affect analysis by different authors, using different modalities. Some authors have used uni-modal signals, while others have used multi-modal signals to achieve the same purpose of analyzing affect. Some of these works are described in the following paragraphs:

Multi-modal datasets

Dataset for Emotion Analysis using Physiological Signal (DEAP), a multi-modal dataset for the analysis of human affective states was presented in [81]. The dataset included 32 participants' electroencephalograms (EEGs) and peripheral physiological signals, as well as 22 participants' frontal face videos. As categorization modalities, electroencephalogram signals, peripheral physiological data, and Multiple Correspondence Analysis (MCA) were used. The participants watched a 40 minutes music video. The three separate binary classification problems presented were low/high arousal, low/high valence, and low/high liking. The test set was classified using a Gaussian Naive Bayes classifier, and feature extraction was done using Fisher's Linear Discriminant. Arousal, valence, and liking, respectively, reported 62%, 57%, and 55% accuracies for EEG signals. The peripheral signals had accuracies of 57%, 62%, and 59%, while MCA signals had accuracies of 65%, 61%, and 67% respectively. The authors concluded that overall classification using MCA features is significantly better than EEG and peripherals.

Emotion recognition using electrocardiogram (ECG) and electroencephalogram (EEG) signals from a wireless low-cost off-the-shelf device (DREAMER), which is a multi-modal database made up of EEG and ECG signals recorded at the time of affect elicitation by audio-visual stimuli, was presented in [103]. Twenty-three individuals provided signals of valence, arousal, and dominance with their self-assessment of affective states following each stimulation. Supervised affect recognition classification tests were carried out in order to use the characteristics acquired by the two modalities, as well as their fusion, to provide baseline classification results for the proposed database. Low/high arousal (calm/excited), low/high valence (unpleasant/pleasant), and low/high dominance (without control/empowered) were described as three different binary classification schemes. The EEG-based features offered the highest accuracy for valence classification, whereas the fusion of the EEG and ECG features produced the highest accuracy for arousal classification (62.32%). The maximum accuracy for dominance was 61.84% for both the EEG-based and fused characteristics. In terms of F1 score, ECG-based features outperformed fused features for valence and arousal (53.05% and 57.98% respectively) whereas fused features outperformed ECG-based features for dominance (53.05% and 57.98%, 61.71%, respectively).

An emotion of Face, Body gesture, Voice, and Physiological signal (emoF-BVP) dataset for multi-modal affect detection was presented in [104]. The dataset includes the face, body gestures, voice, and physiological signals (FBVP). The dataset was made by recording actors exhibiting various emotional expressions. The dataset comprises audio and visual sequences of different intensities of 23 different emotional expressions. Feature learning extensions of well-known deep learning algorithms were used to examine the performance accuracies for emotion recognition in uni-modal, bi-modal, and multi-modal contexts in order to show the application of the emoFBVP database. For multi-modal emotion recognition, the fundamental emotional expressions such as anger, happiness, sadness, contempt, fear, surprise, and neutrality were examined. Three categories of data were created using information from the dataset's metadata: ideal data (all three evaluators agreed on the emotion label), non-ideal data (most evaluators agreed on the emotion label), and

mixed data (a combination of ideal and non-ideal data).

MAHNOB-HCI, a dataset that was recorded in reaction to stimuli with the aim of establishing emotion recognition and implicit tagging study, was presented in [81]. Through the use of a multi-modal system setup, facial videos, audiovisuals, eye-gaze data, and the physiological signals of the central nervous system were all recorded. Twenty people from different cultural backgrounds, including 11 men and 16 women, took part in the trials. A gaze tracker, cameras, and microphones were utilized to record the participant's movements throughout the experiment. A Biosemi Active II device was also used to capture the participant's physiological reactions. Many research laboratories all around the world have employed the Biosemi active II system to capture physiological reactions. The system connects to the rest of the system through a fibre optic (galvanic isolation) and is powered by a battery. Before the experiment, the participant's body was fitted with physiological signal sensors such as EEG sensors, ECG sensors, sensors on the fingers, skin temperature sensor, and a respiration belt around the chest. The participant was asked to calibrate the gaze tracker by following red circles on the screen. The videos and bodily responses recorded were segmented and stored in a database, and made available to the academic community.

Facial Signal Datasets

Another form of modality that was used in affect analysis is the facial database. One of the most significant components in human communication that helps one to grasp what the other person is attempting to communicate is facial cues. Only one-third of the message is understood vocally, and the other two-thirds are communicated through non-verbal means [105]. AffectNet is a database of face emotions gathered, annotated, and produced by Mollahosseini et al. in [106]. The AffectNet database contains over 1,000,000 images gathered from the internet. To obtain the information, three major search engines were queried in six different languages using 1,250 emotion-related terms. The presence of seven basic facial emotions, including valence and arousal intensity, was manually annotated in a reasonable proportion of the facial images. Twelve human experts annotated 450,000 photographs in dimensional (valence and arousal) and category models, and images with occlusion on the face were tagged. The level of agreement among human labellers on a subset of AffectNet revealed that it is difficult to recognize expressions and forecast valence and arousal in the wild. The two annotators disagreed on the value of valence and arousal (RMSE = 0.34 and 0.36, respectively), but they agreed on 60.7% of the category of facial expressions.

Denver Intensity of Spontaneous Facial Action (DISFA) dataset which consists of video clips that are intended to elicit spontaneous emotional expressions was presented in [107]. Twenty-seven participants(12 women and 15 men) with ages ranging from 18 to 50 years old took part in the study. Three Asians, twenty-one Euro-Americans, two Hispanics, and one African-American were among the participants. Participants watched a 4-minute (242-second) video clip designed to generate spontaneous AUs in reaction to videos eliciting a variety of facial expressions of emotion. The video was made up of nine portions, most of which were obtained from

YouTube. Participants watched the film while seated in front of a video screen and stereo cameras. Their facial behavior was captured using a high-resolution (1,024 768 pixel) camera. A total of 4,845 video frames were recorded for each participant. On a 0 (not present) to 5 (highest intensity) ordinal scale, AU intensity was coded for each video frame. The number of events and frames at each intensity level for each AU was recorded. An AU's continuous occurrence from its onset (start frame) until its offset is referred to as an event. Authors of DISFA also presented DISFA+ (extended DISFA) database in [18]. DISFA+ was introduced and discussed along with its use to compare and analyze the temporal patterns and dynamic properties of posed and spontaneous facial expressions. DISFA+ contains the following features:

1. a large collection of data on the posed and spontaneous facial expressions of a group of people,
2. manually labelled frame-based annotations of the 5-level intensity of twelve Facial Action Coding System (FACS), and
3. metadata (i.e. facial landmark points in addition to each individual's self-report for each posed facial expression).

DISFA+ dataset contains videos and AU annotations of the posed and natural facial expressions of nine subjects from the DISFA dataset. The smallest detectable change in facial movement is known as an "action unit" (AU) and can be displayed singly or in groups to represent facial expressions [108]. A High Definition (HD) camera with 1,280 × 720-pixel resolution with a frame rate of 20 frames per second, was used to record the facial responses of the participants. Participants mimicked each facial action a few times, and after each trial, participants scored each mimicked facial expression on a scale of [0-10] by answering two questions: 1) How tough was each facial expression for you to create? 2) Could you accurately imitate/pose the facial expression? The authors selected the best trials from each person with the greatest score. These trials were then FACS coded, with each frame's intensity designated. DISFA+ dataset is publicly available for research purposes.

A comprehensively labelled dataset of naturalistic and spontaneous facial expressions collected in the wild known as Affectiva-MIT Facial Expression Dataset (AM-FED) was presented in [109]. The dataset consists of 242 facial videos of 168,359 sequence frames. The facial videos involved online viewers watching some amusing commercials and being filmed in the course of watching using their webcam. The dataset was labeled by 1) frame-by-frame labels indicating the presence of 10 symmetrical FACS action units, 4 asymmetric (unilateral) FACS action units, 2 head movements, general expressiveness, and gender; 2) the location of 22 automatically detected landmark points; 3) self-report responses indicating familiarity with, liking, and desire to watch the stimuli videos again; and 4) baseline performance of detection algorithms on this dataset. Extended Affectiva-MIT Facial Expression Dataset (AMFED+), which was presented by the same authors in [110], is an expanded dataset of naturalistic facial response videos that were gathered in common contexts. One thousand and forty-four (1,044) videos totalling 263,705 sequence frames were manually coded for facial action units in 545 of the 1,044 videos that made up the dataset. The labels matched those in [110].

Facial Emotion Recognition-2013 (FER-2013) was presented in [111]. The dataset was created by looking for images of faces that matched a list of 184 emotion-related keywords, including "blissful," "enraged," and other terms, using the Google image search Application Programming Interface (API) [111]. In order to construct over 600 strings that were used as facial image search queries, these terms were combined with terms like gender, age, and ethnicity. Using OpenCV face recognition [112], bounding boxes were created around each face in the collected images. While filtering out some duplicates, human labellers rejected photographs that were improperly classified. After being approved, the images were cropped, resized to 48x48 pixels, and made grayscale. The fine-grained emotion keywords were divided into seven main categories by the authors after they developed a sample of the images. The final dataset contains 4,953 "Anger" images, 547 "Disgust" images, 5,121 "Fear" images, 8989 "Happiness" images, 6,077 "Sadness" images, 4,002 "Surprise" images, and 6,198 "Neutral" images. Because of how images were collected, FER-2013 may have contained label errors. However, human accuracy on FER-2013 reported was 65.5%.

DAiSEE was a dataset developed for the analysis of affective states [19]. Video recordings of subjects in an e-learning environment were annotated with crowdsourced learning-related affective state labels for engagement, confusion, boredom, and frustration. The dataset captures "in the wild" conditions that are common in the real world. A full HD web camera (1920x1080, 30 fps, focal length 3.6mm, 78 fields of view) installed on a computer to replicate real-life conditions, focusing on learner users watching movies was employed. A special program was built to deliver a subject with two distinct videos, each lasting 20 minutes, that are both educational and entertaining in order to establish a virtual e-learning environment. These were used to record both concentrated and unfocused states, allowing for variations in user involvement levels that occur naturally. Individuals were given the option of scrolling through the movies to simulate unrestricted situations. The dataset contains 112 participants between the ages of 18 and 30, all of whom were actively enrolled learners. The subjects were Asian in race, with 32 females and 80 males. A total of 12583 video samples, each lasting 10 seconds were gathered. The authors experimented with both temporal and static models. CNNs were used in all the models. The models used to benchmark each emotional state individually included the InceptionNet frame level, InceptionNet video level, C3D training, C3D fine-tuning and Long Term Recurrent Convolutional Neural Network (LRCN) models. All the classification models were tested on Top-1 accuracy and each three times to avoid any randomization bias. It was observed that LRCN outperforms all other models in most cases, reporting accuracies of 53.7%, 57.9%,72.3%, and 73.5% on boredom, engagement, confusion and frustration respectively. DAiSEE dataset is also publicly available for research purposes.

Speech/Voice Datasets

Another form of modality through which affect can be analyzed is speech or voice. Speech portrays affective information via the explicit linguistic message, that is, what was said and the implicit features of the expression, that is, how it was said. A

database for speech emotion known as Emotion Database (EMO-DB) was described in [113]. EMO-DB contains utterances of 10 speakers, made up of 5 males and 5 females. joy/Happiness, anger/fury, neutral tone, sadness, fear/anxiety, boredom and disgust are among the 804 data points in the Emo-DB collected from the 10 participants. The recordings were made with high-quality recording equipment in an anechoic environment. Electro-glottograms were recorded in addition to the sound. About 800 sentences made up the spoken material (seven emotions, ten performers, ten sentences plus some alternate versions). A perception test was used to assess the database's overall recognizability and naturalness of emotions. Utterances that were recognized by more than 80% of listeners and rated natural by more than 60% of them were phonetically marked in a restricted transcription with particular markers for voice quality, phonatory, and morphological features.

An annotated and standardized corpus in the Spoken Dialogue System (SDS) area was introduced in [114]. The database includes non-acted American English utterances that were acquired from the Let's Go Bus information system (LEGO) at Carnegie Mellon University. Nine thousand and eighty-three (9,083) system-user label exchanges and 347 chats were included. The elements of Automatic Speech Recognition (ASR), Spoken Language Understanding (SLU), and Dialogue Manager (DM) have been combined to automatically and semi-automatically generate a total of 46 parameters. Each spoken user utterance has been given a label from the collection of emotions, including garbage, non-angry, slightly angry, and extremely angry. Additionally, three raters manually annotated Interaction Quality (IQ) at the level of exchanges, yielding a Kappa value of 0.54. The "IQ" value, which ranges from 1 to 5, describes the level of involvement up until each user-system contact. The corpus supplied here is intended to serve as a standardized basis for task success prediction, dialogue quality estimate, and emotion identification tasks, allowing for comparison between different approaches in these disciplines.

Utsunomiya University Spoken Dialogue Database for para-Linguistic Information Studies (UUDB) database was presented by the authors in [115]. The UUDB database contains spontaneous Japanese human-human dialogue. A total of 4,737 utterances were produced through task-oriented discussion between seven pairs of speakers (12 females and 2 males). Three annotators assigned emotional labels to each utterance based on a five-dimensional emotional scale (interest, credibility, dominance, arousal and pleasantness). Only pleasantness (or assessment) and the arousal axis were utilised in their work. The emotional labels were then allocated to the respective quadrant (clockwise, starting in the positive quadrant, assuming arousal as abscissa): happy-exciting, angry-anxious, sad-bored, and relaxed-serene.

Written Text Datasets

Written text is a modality that can be utilized to recognize affect-related emotions. Written texts can be used to detect emotions. Semantic Evaluation (SemEval-2020) dataset was proposed in [116]. The dataset comprised news headlines that had been edited to make them humorous for the SemEval challenge. This dataset was annotated using Amazon Mechanical Turk, where qualified human employees modified headlines to make them humorous, and a separate group of qualified human

judges judged the quality of humour in these headlines on a 0-3 funniness scale. This method of quantifying humour paves the way for the creation of software that can automatically estimate the level of humour in a text. Their task was divided into two parts: Subtask 1: On a $0-3$ humour scale, rate the amusement of a modified headline. Subtask 2: Determines which of the two altered versions of the identical headline is funnier. Subtask 1 attracted 48 participants, while Subtask 2 attracted 31. The top performer in both subtasks reported an accuracy of 67.43 % (a 17.93% improvement over BASELINE which was 49.5%) for Subtask 2, and an RMSE of 0.49725 (a gain of 0.2% over BASELINE) for Subtask 1.

EMOBANK [117] is one of the datasets for assessing written language's emotional content. The ten thousand (10,000) English sentences in this dataset were taken from news headlines, essays, blogs, newspapers, fiction, letters, and travel guides written by people of different genders and were annotated with dimensional emotion data using the Valence- Arousal- Dominance (VAD) representation format [117]. A rudimentary emotion model developed by Ekman was used to annotate the dataset. EMOBANK's bi-perspectival and bi-representational designs were exceptional [117]. In addition, a section of the corpus complements dimensional VAD annotations with categorical ones based on basic emotions. The authors discriminated between the writer's and reader's emotions. The reader's perspective was clearly superior in terms of IAA and rating intensity, and they were able to translate between dimensional and category formats with almost human performance.

A dataset known as Workshop on Computational Approaches to Subjectivity Sentiment, and Social Media Analysis (WASSA-2017) was designed to detect emotions in tweets [118]. WASSA-2017 was labelled for four discrete emotions, which include anger, fear, sadness, and joy. The aim of creating the dataset was to assist in the understanding of the levels of emotions displayed in human language through writing. Authors of [119] built a dataset called Daily Dialogue. The Daily Dialogue dataset was built from crawling dialogues from day-to-day conversations. The dataset is made up of a total of 13,118 sentences labelled neutral, disgust, happiness, surprise, fear, anger and sadness emotions.

Body Language Datasets

Posture can give some information that is sometimes not available from the usual measures, such as the face and speech. The affective state of a person can be detected over long distances with posture [120]. However, capturing features of the other modalities at the same distance is difficult. This gave rise to more studies in affect analysis using body language or posture over the last decade. An Unsupervised Ranking using Magnetic Properties and Correlation Coefficient (URMC) dataset for predicting affect from body language was proposed in [121]. The framework was used to forecast mental health symptoms. The dataset contains 144 30-second video segments cut from 12 20-minute videos. Each of the 14 body language labels, 24+1 emotion labels, and 24+1 symptom labels has numerous labels on it. Forty-eight of the 144 movies were chosen for training and received frame-level body language descriptions. In addition, 48 videos were chosen for validation and 48 videos for testing. The split is based on a cross-subject setting.

2.5 Techniques used for Affect Analysis

Like any other framework, affect analysis include stages such as pre-processing, feature extraction, and classification. In sections 2.5.1, 2.5.2 and 2.5.3 we review the methods and techniques used in literature as part of the above-listed stages.

2.5.1 Pre-processing Methods

To extract significant features for image recognition, image preprocessing is employed to improve the image data. To focus on the area of interest in the image, image cropping is performed to remove the image's unimportant sections. In some circumstances, image morphological techniques can be used. The resulting image's additional intensity values are added via image filtering. To remove noise or other little unnecessary data from the image, smoothing techniques are applied [122]. Preprocessing is typically used to enhance, restore, and compress image data as well as fix radiometric or geometric problems [123]. Data preprocessing is a phase in data mining and data analysis that converts raw data into a format that computers and machine learning can comprehend and evaluate. Data preprocessing is the process of modifying or encoding data in order to make it rapidly and easily decoded by a machine. In order to improve image features by removing unnecessary falsification, attention to image pre-processing is necessary for the best computer vision results. For a model to make an accurate prediction, the algorithm must be able to quickly examine the data's features. In our machine-learning approach, features serve as discrete, independent variables that are input. These variables can be thought of as representations or features that provide the data context and help models predict classification and labelling. Image preprocessing may enhance the accuracy of feature extraction and the results of image analysis, as mentioned by [124]. The following are a few image preprocessing techniques:

Face Detection and Localization

Face detection is a type of application that is part of computer vision technology. Face detection makes use of a number of classifiers, which are algorithms that identify whether an image contains a face. In the study by [125], the Haar Cascades approach was used to determine whether the input image contained a face. Before using the feature descriptors, the authors in [126] and [127] employed the Viola-Jones face identification method to locate a face in an image. To extract the angled local directed pattern for texture feature analysis, Sang, et al.[128] detected the face using Haar-like features. The facial image preprocessing with a multi-task CNN face detector for face detection, normalization, cropping, and scaling to 120 $\times$ 120 pixels was part of deep neural network emotion recognition framework proposed in [129].

Facial Parts/ Landmark Detection

Another preprocessing step that could be applied to facial images before feature extraction is to detect the facial parts or landmarks. Facial landmarks are the

significant characteristics of a face, and these include, for example, the centre of the eye and the tip of the nose [130]. Facial landmark detection is described in computer vision as the task of finding and tracking important facial landmarks[130]. For example; let I be the input image, which is represented as a 3-dimensional tensor with the dimensions $W \times H \times C$, where W, H, and C are the width, height, and a number of colour channels of an image, respectively. Three channels are normally utilized for colour images, one for each of the red, green, and blue colours. The challenge of face landmark detection is to find a function $\Phi : I \longrightarrow Y$ that predicts a landmark matrix $\hat{Y} \epsilon R^{N_L \times 2}$ from the input image I. N_L is the number of facial landmarks, $\hat{Y}_{i1} \epsilon [0; W]$ representing X coordinates and $\hat{Y}_{i2} \epsilon [0; H]$ representing Y coordinate of the i^{th} landmark [13].

The correct detection of the landmark points is the foundation of many face analysis methods, and the localization of the fiducial facial key points is typically a crucial step according to [13]. These key points are shown in Figure 2.13. Six sub-regions were identified in [131] which include; the right cheek area, forehead, left cheek area, nose, eyes and mouth. The authors in [132, 133] used the Viola-Jones detection algorithm which used Haar features to locate the facial part. For the purpose of an emotion recognition system, a Haar Cascades approach to detect faces and crop the face into the important regions of the lips and eyes was presented in [125].

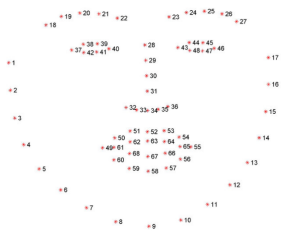


Figure 2.13: 68 Facial landmark Point [13]

2.5.2 Feature Extraction Techniques

As it has been already noted, the human face is an informative representation such that a facial image is represented by massive, intricate, and unintelligible numerical data. The term "feature" refers to the subject information contained in the image data, while "feature extraction" refers to the accurate extraction of usable features from the image data. As feature extraction removes redundant variables from the data to avoid computational complexity, overfitting, and non-generality of the feature model, feature extraction typically results in a downsizing or dimensional reduction of the dataset.

Feature extraction techniques are used to obtain features that can be used to categorize and identify images. Dimensionality reduction is useful as a pre-processing step for machine learning because it decreases unneeded and redundant data, boosts learning accuracy, and improves result interpretability. Dimensionality reduction is a crucial pre-processing step that improves the accuracy and efficiency of data analysis computations. It also has various advantages, including removing irrelevant

and redundant patterns from the dataset, which consequently reduces the time and memory required to process such data. It is a crucial pre-processing step that improves the accuracy and efficiency of data mining computations. Advantages of dimensionality reduction include removing irrelevant and redundant patterns from the dataset to reduce the time and memory required to process such data [134, 135]. An example of a dimensionality reduction process is feature extraction.

Feature extraction is a comprehensive approach, which entails attempting to transform the input space into a low-dimensional subspace while retaining the majority of the crucial data [136]. The goal of feature extraction and selection approaches is to increase performance such as estimated accuracy, visualization, and comprehensibility of learned knowledge [137]. The extraction of features is a critical step in the emotion classification process. The accuracy of emotion categorization is directly influenced by the quality of the feature extraction process [138].

Feature extraction is an important issue in machine learning and pattern recognition, and various algorithms have been presented, especially for affect analysis. The appearance-based, geometric-based, learning features-based, and hybrid-based methods are some of the common feature extraction techniques for an affect analysis task. Geometry-based and appearance-based methods of feature extraction are categorized as handcrafted feature models. Different feature descriptor algorithms are used in each of these techniques. Some of these techniques are highlighted and described in the following subsections.

Hand Crafted Feature Extraction Technique: Appearance and Geometric Based

Facial features may be categorized as either global or local features in appearance-based models. This technique removes changes in the face image by combining an image filter or filter bank with either the entire image or a specific area of interest in the image [139]. Global feature descriptor approaches are used to translate the image features into a single multidimensional feature vector of color, shape, or texture. The N-dimensional feature vectors are created by the number of interest points, but the local feature descriptor technique emphasizes interest points (key points) more. The following are some examples of appearance-based feature extraction methods for analyzing affective states.

Gabor Wavelet: This feature extraction method was named after Denis Gabor in 1946[140]. Image analysis using Gabor detects a region in an image with a given frequency in a specific direction. Gabor was suitable for image texture representation and discrimination due to the frequency and orientation descriptions. A Gabor filter is a function that captures the appropriate frequency spectrum in a spatial domain by amplitude modulating a sinusoid with a Gaussian function. Gabor Wavelet is represented mathematically as:

$$G(X,Y) = \frac{1}{2\pi S_x S_y} exp\left[-\frac{1}{2}\left(\frac{x'^2}{s_{\frac{x}{y}}^2}\right)+\left(\frac{y'^2}{s_{\frac{x}{y}}^2}\right)\right] exp\left[j\frac{2\pi X'}{\lambda}\right] \tag{2.1}$$

Assuming (x, y) to be the pixel position in the spatial domain, α to be the wavelength in pixel, θ to be the orientation of the Gabor filter and Sx, Sy to be the standard deviation along the x and y. $X' = x\cos\theta + y\sin\theta$ and $Y' = -\sin\theta + y\cos\theta$

Histogram of Oriented Gradient (HOG): HOG is a feature descriptor used in a variety of domains, where object description is important through shape and appearance [141]. The HOG descriptor was motivated by the distribution of intensity gradients within an image, which can represent local object appearance and form as well as related edge directions [142]. As a feature descriptor for image region description, HOG is widely used in object detection. HOG begins by separating an image into blocks, which are then further divided into cells. Because of the overlapping blocks, the cell became a subcell of multiple blocks, and the vertical and horizontal gradients were produced for each cell's pixel [143]. If the vertical gradient is $G_y(Y, X)$ and the horizontal gradient is $G_x(Y, X)$, then the magnitudes of the gradients are derived as indicated in 2.2 and 2.3, respectively.

$$G(X, Y) = \sqrt{G_x(Y, X)^2 + G_y(X, Y)^2} \tag{2.2}$$

$$\theta(Y, X) = \arctan \frac{G_y(Y, X)^2}{G_x(Y, Z)} \tag{2.3}$$

HOG features is generated for each cell. The total number of bins containing the description is the sum of these histograms. As different images have varying contrast, contrast normalisation is required to improve performance. This normalisation results in insensitivity to variations in lighting and shadows [144].

Principal Component Analysis (PCA): Image data is a high-dimensional dataset, making pattern extraction difficult. PCA, on the other hand, can achieve pattern recognition and a degree of data variability. PCA projects the relation between variables in high-dimensional data into a more tractable lower-dimensional version without losing much information [145]. PCA has a tendency to select an axis system in data that points to the largest correlation in the given data. By using only the relevant Eigenfaces responsible for apparent variability, image data reconstruction results in substantial dimensionality reduction [146].

Learned Features

Convolutional Neural Networks: CNNs has revolutionized the approach to gathering learned features [147]. Without using feature extraction mathematical models, CNNs learned the direct representative features from the input. Breuer et al. [148] used deep learning visualization tools to see what kind of feature CNN employed for classification, and discovered that the features at the low level matched low-level Gabor filters. The main issue with applying learned-based features to affect analysis is a lack of sufficient data for a network to learn, which leads to overfitting.

Transfer learning: Transfer learning is another high-performing learning-based feature strategy, and it is especially effective in affect analysis, where there is minimal

data for model training. Transfer learning refers to a model's capacity to learn patterns from data that it was not previously trained on [149]. Generally, there are two forms of transfer learning applicable to deep learning for computer vision, that were identified in [150, 20]. The first is to treat networks as feature extractors, and the second is to remove fully connected layers from an existing network while placing a new FC layer set on top of the CNN, to fine-tune the weights to recognize object classes. over the last decade, researchers have proposed several pretrained CNN models to enhance the system's performance. Some of these include the Residual Network (ResNet) [151], Xception [152], Visual Geometry Group (VGG)[153] which includes VGG-16, VGG-19, Inception ResNetV2 [154], InceptionV3 [155], DensNet [156], MobileNet [157]etc., all trained on Imagenet datasets. ImageNet is an image database arranged according to the WordNet hierarchy (currently only nouns), with thousands of images representing each node of the hierarchy [158]. The project has made major contributions to the disciplines of deep learning and computer vision. Researchers can access the data for free and for academic purposes. Instead of training a classifier from scratch, pre-trained networks have been used for both feature extraction and classification and so far they have been proven to be more efficient than training a classifier from scratch. Due to its resistance to illumination, rotation, translation, and head pose learned features have demonstrated promising outcomes, particularly in affects analysis and FER [144]. Transfer learning models outperform typical machine learning models in terms of performance. This is because the models that employ knowledge (features, weights, etc.) from previously trained models already know what the characteristics are. The process is faster than training neural networks from the ground up.

Three pre-trained CNNs, namely, VGG-16, ResNet-50, and SE-ResNet50 was presented in [159]. ResNet-50 reported the highest validation accuracy of 99.47%. Rajpal et al.[160] explored the strength of transfer learning using ResNet-50 for training a network on a set of preprocessed images to classify COVID-19 from chest x-ray images. Their proposed model generated an overall classification accuracy of 97%. Work by Al-Haija et al. in [161] was used to train and classify the BreakHis dataset using transfer learning technique of ResNet-50 CNN that was pretrained on ImageNet. The simulation results showed that their model outperformed other compared models trained on the same dataset with an exceptional classification accuracy of 99%. Several deep learning models, including Inception-V3, VGG19, and ResNet-50, were evaluated and compared to determine the best predictive classification model for real-time engagement detection in [162]. This is accomplished by analyzing the learners' facial expressions throughout the online learning session to classify their emotions. On benchmarked datasets, the proposed system reported accuracies of 89.11%, 90.14%, and 92.32% for Inception-V3, VGG19, and ResNet-50, respectively. In real-time learning scenarios, ResNet-50 outperforms all others with an accuracy of 92.32% for facial emotion classification.

VGG-19 is another CNN-pretrained network that has been used for feature extraction in various research studies and has recorded remarkable results. Some of these include work by Smith et al in [163] who tested the effects of changes in various design schemes and training parameters on prediction accuracy using both VGG19 and VGGFace pretrained models. Input standardization, data augmentation, and

label distribution age encoding were all compared as training techniques. Finally, a deep CNN hierarchy was tested, which first classified subjects by gender before using separate male and female age models to predict age. The gender recognition accuracy was 98.7%. A transfer learning solution for compound expression recognition using a fine-tuning structure combined with a VGG-19 pre-trained network for facial expression recognition [164]. Deep facial expression features were extracted using the VGG-19 pre-trained network, and the facial expression category was determined using the fine-tuning network. According to the experimental findings, using VGG-19 for feature extraction increased the accuracy of the most advanced deep learning technique by 5.31%.

Hybrid-Based Features

The hybrid properties allow for the research question that aims to establish the best approach to combine image features to attain maximum performance. An hybrid method that used LBP and HOG characteristics derived from the CK+ and JAFFE databases was proposed in [165]. PCA was used to minimize the dimension of the extracted features. The fusion was permuted on various classifiers. It was noted that a Softmax classifier's features yielded 98.3% on the JAFFE database, and 90% on CK+, which signified that proper hybrid features could vastly improve a system's performance.

2.5.3 Classification Methods

Despite the widespread success of affect analysis based on the extraction of handmade features, researchers have turned to deep learning for classification in the last decade because of its automatic recognition capability. Some of the algorithms used in deep learning are organized in a hierarchy of increasing complexity and abstraction. Each algorithm transforms its input nonlinearly and outputs a statistical model based on the features it has learned. This procedure is iterative, up until a discernible degree of accuracy is attained. CNNs and RNNs are the most widely utilized deep learning neural networks in computer vision. Some others include Capsule Networks, Cascaded Networks, Multitask Networks, and Generative Adversarial Networks (GAN). In this regard, this study will describe some of the affect analysis research that demonstrates proposed deep learning strategies for enhanced detection.

Convolutional Neural Network

CNN, whose design originated from ANNs [166], is one of the most widely used deep learning architectures for computer vision and image processing domains. It uses end-to-end learning, which is carried out through a hierarchy of layers. Each CNN layer generates representative features that range from the image's basic elements to a more abstract idea. There is no need for any pre-calculated feature extraction techniques because CNN models can extract their representative features on their own. Its outstanding performance contributes significantly to its popularity. The convolution layer, pooling layer, dense layer, and fully connected layer are the

four primary layers that make up CNN architecture [166]. With the advancement of technology and a reduction in CNN limits to FER/affective state analysis, its application continues to rise. There have been many studies on affect analysis using CNN as a basic classifier in various configurations.

Song et al. [167] introduced a DGCNN-based multichannel EEG emotion classification method. The suggested EEG emotion recognition approach was based on the premise of utilizing a graph to characterize multichannel EEG characteristics and then using this model to perform EEG emotion categorization. The proposed DGCNN method, in contrast to conventional GCN methods, uses a neural network to dynamically learn the intrinsic relationship between various EEG channels. To benefit from more discriminative EEG feature extraction, this is represented as an adjacency matrix. Following that, more discriminative qualities are taught using the learned adjacency matrix to improve EEG emotion recognition. With an average recognition accuracy of 90.4% for subject-dependent experiments and 79.95% for subject-independent cross-validation experiments on the SEED database, as well as an average accuracy of 86.23%, 84.54%, and 85.02% for valence, arousal, and dominance classifications on the DRE, the experimental results show that the proposed method outperforms state-of-the-art methods.

Recurrent Neural Network

Feed-forward neural networks (FNNs) with hidden memory nodes are referred to as Recurrent Neural Networks. The word "recurrent" comes from the way that RNNs works, where the output of the current input is dependent on the outcomes of processing the previous input(s). RNN is suited for a variety of sequence-related applications, including joined handwriting, voice and speech recognition, Natural Language Processing (NLP), and video processing owing to the hidden nodes. Since RNN was introduced, effects analysis could be applied in dynamic or spatio-temporal environments [168]. However, gradient vanishing and exploding is the fundamental issue with RNN. The majority of studies that modelled effects analysis in a spatiotemporal setting employed LSTM, a modified RNN that keeps track of previous data in memory and circumvents gradient vanishing difficulties. LSTM has an advantage over RNN in that it can handle information in memory for a longer amount of time than RNN (retaining information for an extended length of time) this is because LSTM has a "memory cell" that can store information for extended periods of time and so could learn longer-term dependencies.

Long-Short-Term Memory

Several researchers have used LSTMs for classification tasks using temporal data. Manessi et al. in [169], for example, provided two techniques for learning long short-term dependencies and graph structure. Waterfall Dynamic-Graph Convolution Network (WD-GCN) and Concatenate Dynamic-Graph Convolution Neural network (CD-GCN) were the two approaches used. These techniques were developed using a combination of LSTM networks and GCN. The model, which consists of a semi-supervised classification of vertex sequences and a supervised classification of graph sequences, was developed to classify dynamic graph-structured databases

using a mix of graph convolutional GCunits and LSTM. In comparison with other baselines like FC+FC, GC+GC, and LSTM-based, the LSTM-based baseline reported the best performance on WD-GCN and CD-GCN, with reported accuracies of 70.1% and 69.3%, respectively. Cheng et al. [170] proposed a deep forest-based technique for multi-channel EEG-based emotion identification. Baseline removal was used to pre-process the raw EEG data after artifact removal. This was done with the baseline signal in mind. Second, 2D frame sequences were created by accounting for the spatial link between channels. The 2D frame sequences were then loaded into a deep forest classification algorithm, which uses the spatial and temporal data in the signals to categorize EEG emotions.. Traditional methods require feature extraction, while the proposed method does not, and the classification model is unaffected by hyper-parameter settings, which considerably reduce the complexity of emotion recognition. Experiments were carried out on two public datasets, namely, DEAP and DREAMER, to verify the practicality of the suggested approach. The average accuracies for valence and arousal on the DEAP database were 97.69 % and 97.53 %, respectively, on the DREAMER database. The average accuracies on DREAMER datasets were 89.03%, 90.41% and 89.89 %, respectively. These findings suggest that the proposed method outperforms existing methods in terms of accuracy. A multi-modal technique for predicting human emotions in the continuous valence dimension based on a feature-level fusion of human facial expressions and EEG data was proposed in [171]. A Recursive Neural Network with Long Short-Term Memory units LSTM-RNN was developed for this purpose. The MAHNOB-HCI data set used to test the performance of the suggested technique reported 0.041 RMSE. The LSTM architecture is applicable in the following fields of artificial intelligence: robot control [172], time series prediction [173], rhythm learning [174], music composition [175], grammar learning [176], handwriting recognition [177], human action recognition [178], sign language translation [179], time series anomaly detection [180], etc. Siami et al. [181, 182] opined that having LSTM in two directions will allow additional training by traversing the input data twice. For this reason, a bidirectional LSTM was proposed in [182]. Experimental results noted that LSTM with BiLSTM and reported that BiLSTM offered better predictions than regular models based on LSTM.

Bidirectional Long Short Term Memory

Long-term dependencies can be learned with the aid of the BiLSTM structure, which also has the potential to improve model accuracy. [181, 182]. The bi-directional LSTM is an advanced LSTM network made up of two LSTMs [183]. The input sequence data are applied to by the first LSTM in the forward direction, while the input sequence's reverse structure is used by the second LSTM. Long-term dependencies can be learned with the help of this kind of LSTM structure. BiLSTM has been used for classification and has reported state-of-the-art results. Joshi et al. in [184] proposed the use of BiLSTM to classify a subject's emotional state of mind using EEG signal and compared the results with a baseline MLP. Their study used both methodologies to analyze EEG data on 14 patients using Intelligent Database for Emotion Analysis (IDEA) database. On IDEA database, subject dependent av-

erage accuracy of 98.5% was reported, while the subject independent accuracy was 88.57%. The publicly accessible SJTU Emotion EEG Dataset (SEED) and DEAP database were also subjected to a new method of modified differential entropy and BiLSTM network to confirm improvement in the accuracy level. The average accuracy of emotion detection was shown to be higher utilizing the BiLSTM network for classification and modified differential entropy (MD-DE) for feature extraction.

2.5.4 Emotion Analysis in E-learning

One of the tools that have been widely used to study the affect of learners in the classroom is emotion. Tang et al.[185] opined that the teacher in a class could estimate the level of comprehension of the lecture by their learners' facial expressions. They further iterated that learners usually smile after comprehending the lecture but form negative expressions when they could not comprehend. Sathik et al.[25], indicated that facial expression ranks high on the mode of nonverbal communication among all the non-verbal modes of communication such as body language, gestures, hands, and facial expressions. Furthermore, facial expression was expressed in their work as playing a crucial role in identifying the emotions and comprehension level of learners in virtual classrooms. Furthermore, they explained that the most meaningful way humans show emotions is through facial expressions. As a result, they proposed a technique for locating learners' expressions in static images of human faces to determine their comprehension level. Their study's purpose was to categorize facial expression states as comprehensible or incomprehensible. They concluded that, along with words, facial expressions are the key source of knowledge, and their findings were used as feedback to the learning management system [25].

In the e-learning-based classroom, one of the significant challenges was the inability to effectively capture the emotions of the learners, which, if captured, would have played an essential role in the teaching and learning activity [14]. The work in [25] stated that capturing the learner's emotions and improving their mood is missing in e-learning. Consequently, there is a need to build a more interactive system that will remove boredom in learners and thereby improve learning outcomes in the e-learning environment. Even though many available systems have emerged over the years to solve this issue, none better analyzes emotions in real time. Therefore, Krithika et al.[14], came up with the facial emotion recognition system. In their work, they devised a learner emotion recognition system to detect learner emotions based on eye and head movements. Their system was able to recognize learners' emotions in real time, and a feedback mechanism was built into it to serve as a source of information for the instructor. The framework of the system is shown in Figure 2.14.

Further work in the e-learning domain by Mohri et al. [186] presented a learner monitoring algorithm using statistical data acquired through tests, quizzes and assignments. John et al. [187], also presented a recommender system in the e-learning domain. This system was developed to help overcome the challenge of learners searching for useful materials for their learning needs. Another example in the e-learning domain is the use of facial emotion recognition. The authors in [12] proposed that it is possible to analyze the learning affect of all the learners in the

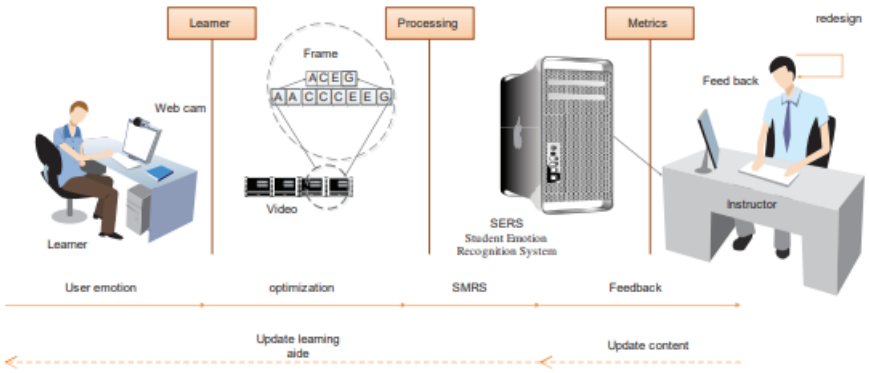


Figure 2.14: Student Emotion Recognition System [14]

class and evaluate the effects of classroom teaching. This was achieved by adopting a stimulus-response mechanism. The learning of facial expressions and attention was used to categorize learners' learning affects, which include curiosity/thinking, comprehension, concern, disregard, and disgust. These facial expressions were used to construct a Students' Learning Affect (SLA) transfer model. Pan et al. [12] described the use of SLA analysis results to auto-evaluate classroom teaching and learning affect. Their study was regarded as a theoretical foundation for other studies on learning facial expressions and learning affects on classroom teaching and learning auto-evaluation. Their work showed that emotions are very significant in identifying how learners learn. Figure 2.15 shows the framework of their model for a classroom automatic evaluation system based on expression recognition learning.

Figure 2.15: Learning Expression Recognition model [12]

A standardized mapping method between face emotions and learning affect was investigated by Zakka et al. in [27]. Their research looked at physical changes in a learner's face to assess facial emotions and then used a mapping mechanism to map emotional states to a learner's learning affect. Using the Facial Emotion Recognition-2013 (FER2013) dataset, a CNN was utilized to classify seven basic face emotions. Their study reported an accuracy of 64.43% on the test dataset.

2.5.5 Affective State Analysis in E-learning

In the domain of e-learning, one significant aspect that can imbibe reasoning is the evaluation of learners' affective states. Analyzing various affective states has an

important role in estimating the learning affect. Each of the affective states carries some information that can be inferred from, as the real state of the individual's mind. Few studies in the area of e-learning have shown that basic emotions are not the most frequent during a short learning session, especially during online learning sessions. Instead of these basic emotions, affective states of happiness, engagement, boredom, anxiety, confusion, frustration, and curiosity were identified to be more prevalent during online learning [29]. The estimation of the engagement of learners within the online environment using facial images was examined in [54]. The work measured the state of engagement of learners during online learning, using the InceptionNet architecture and achieved accuracies of 36.5%, 47.1%, 70.3%, and 78.3% of engagement levels for boredom, engagement, confusion, and frustration respectively.

The application of a deep learning technique on FaceNet embeddings and facial landmark points for academic emotion identification using the DAiSEE dataset was described by Leong et al. in [55]. The emotional states of engagement, boredom, frustration, and confusion were labeled on the dataset. Their work was able to detect instances of boredom and frustration experienced by learners in an online class, with an accuracy of 52.1% and 70.67% on both affective state expressions, respectively.

In an online learning environment, the idea of automatically detecting a learner's affective state was considered by Anne et al. in [188]. A model to identify affective states in texts written by learners was put forth. Machine learning methods and the Kaggle data repository's ISEAR and Twitter data were used. Tokenization, noise removal, and segmentation are methods that were utilized for text preparation. Count vectors and term frequency-inverse document frequency were used to extract features. Various machine learning techniques for the classification of learners' affective states were compared. Their results indicated that the accuracy of Linear SVM utilizing Count Vectors was 79%. They concluded that one might automatically extract the learners' affective states from a text when they are interacting with an e-learning environment. They further concluded that this is expected to aid in comprehending the wants of the learners and in improving adaptability.

Subsequently, a framework for monitoring learners' learning states (LS) as they view various pieces of knowledge-based content and the examination of learners' affective states was introduced by Juan et al. in [189]. Four learning states: engaged, disengaged, confused, and bored were identified. Their model was based on CNNs and Inverted Residual structure. According to their findings, there were sudden changes in the learners' LS, which suggests that this information is closely tied to how learners learn. On the Affectnet, CK+, and DAiSEE datasets, their model reported test accuracy of 73%, 91%, and 61%, respectively.

Learners learning affect was categorised through a video that captured facial expressions while learning by Pan et al. in [12]. Five significant emotions based on the learners' learning expression recognition system, which include joyful, surprised, confused, focused, and distracted were identified. The identified emotions were used to auto-evaluate the learners' learning affect. The work showed that emotions are very significant in identifying how learners learn. A method for locating learners' facial expressions in static images of human faces to identify their comprehension state was presented by Sathik et al. in [25]. The intention of their work was to categorize facial expression states into comprehensible and incomprehensible

categories. Aside from words, it was concluded that facial expression is the most essential source of information. Therefore, facial expression information was used to provide feedback to the learning management system. Using facial images, Dewan et al. [54] also focused on estimating learners' participation in an online context. Using the InceptionNet architecture, these authors were able to accurately quantify the state of learners' engagement throughout online learning. Accuracies of 36.5%, 47.1%, 70.3%, and 78.3% were reported for boredom, engagement, confusion and frustration emotions, respectively.

The use of a deep learning algorithm for FaceNet embeddings and facial landmark points for academic emotion detection using the DAiSEE dataset was proposed by Leong et al. [55]. Engagement, boredom, frustration, and confusion were all identified as emotional states in the dataset. Their work was able to detect boredom and frustration in an online class with an accuracy of 52.1% and 70.67%, respectively.

Table 2.1 provides a summary of some datasets, authors and the year of publication, the classification model used, the affect being modelled, and the accuracies reported.

2.6 Feedback Mechanism

Feedback is a reaction or information that occurs as a result of an individual's or group's actions or behaviour. Positive and negative feedback is essential in a learning and development context [190]. Feedback fosters engagement and interactivity while allowing learners to take ownership of their learning. The support of learners in meeting learning goals and increasing self-regulation skills are revealed through feedback. Providing learners and teachers with information about their learning outcomes is an important part of the teaching and learning process. The learning process is incomplete without feedback as it allows learners or teachers to evaluate their performance in relation to learning objectives to help them meet or exceed their objectives. Any response to a learner's performance or behaviour is referred to as feedback and can be expressed verbally, in writing, or through gestures [191]. The goal of feedback in the assessment and learning process is to help learners improve their grades and to help teachers improve their teaching approach. Feedback is described as an important part of learning structure and learners' support in accomplishing learning goals and increasing self-regulation abilities is revealed through feedback [192]. Feedback can generally be provided for a variety of reasons and in a variety of formats. Feedback can be given as a single entity, for example, informal feedback in class on a learner's understanding of a concept or as a collection of entities, for example, formal, formative, peer feedback in stage one of an assessment task [193]. Each has a role to play in promoting and maximizing learners' learning.

Automatic feedback is a type of feedback generated by artificial intelligence-based software and delivered to learners or teachers upon completion of a teaching and learning task [194]. Automatic feedback offers process-related feedback, can help learners better comprehend their present level of knowledge, and can make recommendations for how and where to improve performance. [195]. It increases learners' performance in activities [192]. The use of automated feedback is a crucial step in enhancing the teaching and learning process. This is because it provides both the teacher and the learner a clear understanding of the issue that arose during

Table 2.1: Recent databases, authors, models and accuracy reported on models for affect analysis

Dataset (Multi-modal)	Signals	Authors and year	Deep learning models	Accuracy reported	Affect modelling
DREAMER	Electroencephalogram (EEG) and electrocardiogram (ECG) signals	Song et al. 2020	DGCNN	86.23%, 84.54% and 85.02%	Valence, arousal and dominance classification
		Cheng et al. 2021	Deep Forest	89.03%, 90.41%, and 89.89%	Valence, arousal and dominance classification
		Tripathi et al. 2021	Capsule network	90.4%	High valence and arousal versus low valence arousal
EmoF-BVP	Face, body gesture, voice and physiological signals.	Ranganathan et al. 2017	Convolutional Deep Belief Network (CDBN)	83.10%	Anger, happy, Sad, disgust, fear, surprise, and neutral
MAHNOB-HCI	Electrocardiogram, Galvanic skin response, skin temperature, respiration volume and EEG signals,	Ben et al. 2017	Support Vector Machine	59.57%, 57.44%	Arousal and Valence
		Zhong et al. 2020	Dynamic Entropy-based Pattern learning (DEPL)	70.25%, 73.27%	Valence and arousal
DEAP	Electroencephalogram (EEG) and Peripheral physiological signals	Verma et al. 2017	Descrete Wave-lenght Transform for feature extraction and MLP, SVM, and KNN for classification	V (51 % and 63 %), A (62 % and 65 %) and D (57 % and 61%)	Valence, Arousal and Dominance
		Khateeb et al. 2021	Time frequency wavelenght + SVM	65.95%	Happy, pleased, relaxed, excited, neutral, calm, distressed, miserable and depressed
		Dhingra et al. 2021	ANN	85.6% and 87%	Valence and Arousal
		Zhuang et al. 2017	SVM	69.1% and 71.99%	Valence and Arousal
		Vaishali et al. 2020	BiLSTM	73% and 75%	Valence and Arousal
		You et al 2020	SAE + LSTM	81% and 74.38%	Valence and Arousal

teaching and learning and how to resolve it. It also helps to understand both teacher and learner performance. Early prediction methods of learner's performance, can be highly helpful to direct learner's learning [196, 197]. Berland et al. in [197] emphasized how crucial it might be to manage and control the problems from the beginning. Automated feedback can be used to gauge a learner's present level of learning and to predict the outcome of that process at its early stages. This could be achieved through effective feedback.

Effective feedback informs learners about their current level of performance and what they need to do to improve [198]. Effective feedback allows the instructor to tailor the learning content to the needs of the learners in this situation. Effective feedback is provided in a suitable setting, focusing on performance rather than an individual. Considering that instructors and learners are geographically and physically separated in online courses, effective feedback is extremely important.

Table 2.2: (Table 2.1 continuation)

Dataset (Facial)	Signals	Authors and Year	Deep learning model	Accuracy reported	Affect modelling
AffectNet	Facial data annotated for the seven basic emotions, including their intensity of valence and arousal	Ichimura et al. 2019	Adaptive DBN	87.4%	Neutral, happy, sad, surprise, fear, disgust, anger and contempt
		Fei et al. 2020	AlexNet +FC6 + KNN	59.8%	Angry, disgust, fear, happy, neutral, sad and surprise
		Hua et al. 2019	Ensemble deep learning model	62.11%	Happy, surprise, neutral, angry, and disgust
DISFA	Facial data, and intensity of facial action units	Akay et al 2021	Stacked CNNs	55%	facial action unit detection
		Hernandez et al. 2021	DeepFN	69.7%	Facial action unit recognition
		Corneanu . et al	Deep Structured Inference Network(DSIN)	56.7%	Facial action unit detection
DISFA+	Videos and action units annotations of posed and spontaneous facial expressions.	Hinduja et al. 2020	CNN	79%	Action unit intensity detection
		Pise et al. 2020	SqueezeNet+TRN	92.7%	Basic emotions of anger, happy, sad, surprise, neutral, disgust, fear
AMFED	Naturalistic and spontaneous facial expression	Vinola et al 2018	Fuzzy logics	86.54%	Intensity of emotions
		Nair et al 2018	Segmentation based fractal texture analysis (SFTA) + Naive Bayes	88.5%	Basic emotions of anger, happy, sad, surprise, neutral, disgust, fear
FER2013	Video dataset of facial action unit, labelled with seven basic emotions	Amal VS. et al. 2021	CNN	73.12%	Basic emotions of the FAC; happy, sad, neutral, surprise, and angry
		Zaka et al. 2020	CNN	64.43%	Basic emotions of anger, happy, sad, surprise, fear, neutral, disgust
		Gowda et al 2021	CNN	65.05%	Basic emotions of fear, anger, neutral, happy, sad, surprise, disgust
DAiSEE	Video sequence of facial emotion expression with labels of four learning related affective states of boredom, engagement, confusion and frustration	Liao et al. 2021	Deep facial Spatiotemporal Network Network (DFSTN)	58.8%	Engagement prediction
		Boughida et al. 2021	ResNet + TCN (Temporal Convolutional Network)	63.9%	Level of engagement of students in an online classroom setting

The following sections review categories of feedback; narrative and Q&A feedback and their significance.

2.6.1 Narrative Reports

A narrative is a written series of events which is a logical outline for a story [199]. The goal of a narrative report is to tell stories about some events that have taken place. One of the advantages of data-to-text narrative report generation is that texts may be easily adapted to certain audiences [200]. NLP tools have been used for the generation of narrative reports of events. Some of these works include a data-to-text module with a highly imbalanced dataset, and a reinforcement learning method with a reconstructor to increase the clinical accuracy of the generated re-

ports [15]. Furthermore, a data augmentation technique for reinforcement learning to train the model on infrequent results was presented. In terms of practical application, a Two-Stage Medical Report Generator (TS-MRGen) was used to generate controllable reports from input photos. TS-MRGen is divided into two stages: picture diagnostic and data-to-text conversion. Radiologists can change the results of the image diagnostic module to influence the reports generated by the data-to-text module. The data-to-text module as well as the complete two-stage approach were tested using two medical datasets. EfficientNet-B4 was employed as the network architecture, which was pre-trained on ImageNet. Results on Japanese Computed Tomography show the highest Clinical Reconstruction (CSR) score of 78.2 was reported and that the reports generated by their model described the findings in the input image correctly. Figure 2.16 presented the framework proposed in [15].

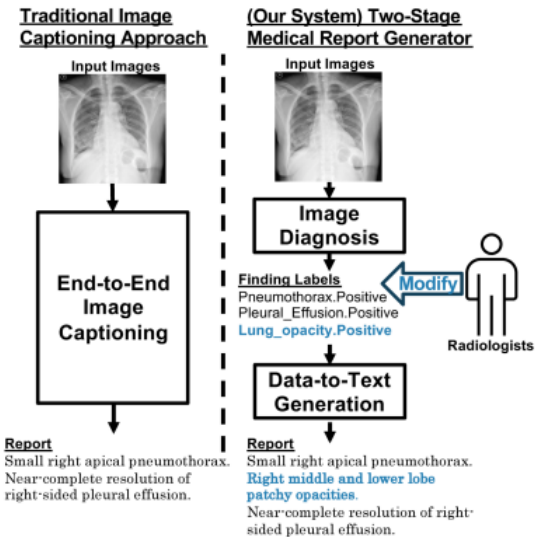


Figure 2.16: An Overview of a Two-Stage Medical Report generator [15]

A lesion-guided explainable few weak-shot medical report generation frameworks that learn the connection between seen and novel classes through visual and semantic feature alignment was presented by Sun et al. in [16]. This was with the goal of generating medical reports for conditions not observed in training. Given a patient's photos, they first extracted global features from each image I using ResNet 101 pretrained on imageNet as the visual extractor of Faster R-CNN. The region proposal network (RPN) then generates lesion region proposals and uses ROI pooling to extract regional visual characteristics. $F^{vis} \cdot F^{vis}$ are divided into three branches: 1. the box regression branch (F_{reg}) detects the location and size of lesion regions; 2. the classification branch (F_{cls}) predicts the category of each region, and 3. the semantic prediction branch (F_prj) projects visual features to semantic features (F_sem) and learns the relationship between seen and novel diseases. Finally, the authors concatenate visual features with their matching word embeddings and feed the concatenated features into Transformer to generate the report. Their framework reported a Bi-Lingual Evaluation Understudy (BLEU) (1 TO 4-gram) result of 0.459,0.311, 0.222, and 0.171, respectively. Their framework is shown in Figure 2.17.

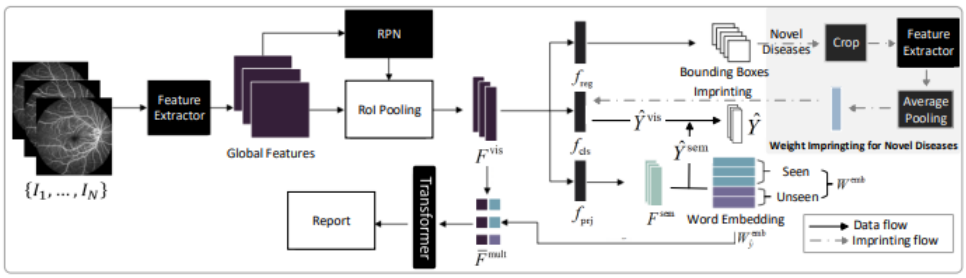


Figure 2.17: Lesion Guided Explainable Few Weak-shot Medical Report Generation Framework [16]

PASS (Personal Automated Soccer test System), a data-to-text system that creates Dutch soccer reports from match information, was introduced by Van et al. in [201]. One of the PASS' innovative features is the way it generates corpus-based texts that are specifically aimed at supporters of one team or the other. This is most obviously seen in the tone of voice used in the reports. Additionally, the system is modular and open source, making it relatively simple for users to use as a test bed for their own. A template and a rule base approach were used in the design of PASS.

2.6.2 Question-Answering System as Feedback

Q&A system is a system designed to answer questions posed in natural language [202]. Q&A systems are information retrieval systems that attempt to respond to questions in natural language by offering answers rather than a simple list of links to documents. It is critical for teachers to establish what they have communicated to their learners during the teaching and learning process. learners exhibit a variety of emotions/ affective states when learning, indicating a link between what they are learning and their understanding of the subject. When there is comprehension during the learning process, there is a connection with the teacher and some emotions/affective states are expressed (voluntarily or involuntarily) by the learners to show the teacher that they have grasped the content. This form of learner understanding is classified as a positive learning affect [32]. On the contrary, lessons can become complex and difficult to comprehend, and learners can disengage from the teacher and express some emotions/ affective states. Teachers may be unaware of the circumstance and, as a result, no learning occurs. This type of result is classified as a negative learning affect [30]. Sometimes, it might be difficult to classify the learning affect as either positive or negative, and this condition is regarded as a neutral learning affect [203]. Unlike in a face-to-face setting, measuring these affect in an e-learning setting is not straightforward. Apart from assessing learners' learning impact through quizzes, tests, examinations, facial expressions, etc., acquiring a real-time understanding of learners while learning in an e-learning setting must be explored further.

For humans, Q&A can be a simple task, but not for a machine. Machines must overcome a variety of obstacles, including lexical gaps, coreference resolution, language ambiguity, and more, in order to answer any question. Machines must have a large training data set and an intelligent architecture to comprehend and store

the text's vital information. [204]. Recent advancements in NLP have unlocked the ability of machines to understand the text and perform different tasks. Some question-answering systems extract data from a source, such as text or an image, to answer specific questions. These sourced systems can be divided into two categories. These are, the open domain, in which questions can be about anything, but are not limited to a specific topic, and the closed domain, where questions can be about anything, but are limited to a specific topic [205]. The closed domain is also called a standard rule-based bot, in which the questions have specified limits, such as relating to a preset source. The closed domain is also called chatbots[205].

Chatbot technology, which is known as a chat robot, is a piece of software that can perform tasks in response to orders or responses given by users during an online conversation [206]. Chatbots are used in a variety of applications, ranging from pure amusement to learning support systems [207]. A chatbot can also be described as a system that interacts with humans and provides solutions to a variety of human-posed questions within its domain [208]. These chatbots operate according to a set of conversational rules, sometimes known as a standard rule base [208]. This implies that a chatbot is a service with which users can interact through an interface, which is a browser dialogue box or an application. One can ask questions by employing choices such as click-based chatbots or rule-based chatbots. One can also type questions in the same way as one would ask a person. Also, a chatbot reacts in the same way as a human would in a discussion [208].

Narekarai et al. in [209], proposed a chatbot system that helps determine the eligibility of students for college admissions. This chatbot system works throughNatural Language Processing (NLP). Front-end coding using Jquery HTML, and CSS was used in the development of GUI and back-end coding was performed using the Java language. The coding shows a rule-based chatbot through which users could interact with the chatbot. A chatbot for university admission using a question-answer system based on a sequence-to-sequence model was developed by Chandra et al. [210]. The model was trained on a dataset consisting of conversations from university admission. The evaluation of the model produced a BLEU score of 41.04 and improved the model to a BLEU of 44.68 using attention mechanism techniques.

An ontology-based chatbot named COVID assistant was proposed in [211]. The chatbot was designed to be a simple-to-use, domain-independent, scalable, dynamic, and intelligent conversational agent. A NLP toolkit was utilized to extract and understand the input, the ontology as a knowledge base, while a finite state machine was used as the chatbot engine. The COVID assistant chatbot, which offers important information about the coronavirus, was created in Python.

From the perspective of conversation design, Illic et al. [212] investigated a deep learning-based dialogue system that creates humorous and sarcastic responses. The basis of their chatbot was a sequence-to-sequence model that was trained on a selected dataset of 3000 question-answering pairs. Furthermore, they used a LSTM based encoder-decoder model in the browser, which allowed users to interact with the chatbot directly. Human rating agents evaluated the system, giving it high marks for personality, sarcasm, and humour. Chatbot systems were examined by Binkis et al. [207] as a technique to assist learners during educational video sessions. They propose a rule-based chatbot system that uses speech recognition and synthesized

responses to enhance immersion and accommodate learners with varying sensory perceptions. These authors explained the design technique they used to create a chatbot system, explained chatbot behaviour from the perspectives of learners and teachers, and defined the structure of the system. The System Usability Scales (SUS) approach was used to validate their system.

A smart chatbot system that can communicate with people and provide them with COVID-19 answers was presented by Amer et al. in [213]. Their model used the pretrained Google Bidirectional Encoder Representation Transformer (BERT) language model to tackle the popular role of Q&A. Two architectural phases for the question-answering task were added on top of BERT. The first step is a text classification technique that uses the BERT to classify text input based on the meaning of the words themselves. The second step entails applying the BERT model as well as the queried domain for answers. Their proposed system was trained and tested on Stanford University's SQuAD V2.0, a well-known question-answering dataset.

2.7 Motivation for the Proposed Study

From the literature review, it is evident that learners' facial expressions and affective states play a crucial role in measuring learners' learning affects. However, most of the studies were limited in their use of estimated emotions and affective states to devise feedback. Feedback in the form of narrative reports and/or answers to the user's (teachers/learners') questions reporting on the learners' behaviour during a learning session and estimated comprehension of the material would be hugely beneficial. Building this form of intelligence into the e-learning domain, particularly in the assessment of learners, should be explored. Therefore, this study intends to contribute to the existing work by proposing a framework to automatically recognize facial emotions and affective states in estimating the learning affect experienced by a learner, generate narratives reports and device Q&A based personalised feedback.

Research indicated that reasoning has been successful in the domain of recognizing objects and their relationship in given scenes [1, 2, 4, 5, 6, 10, 3, 65, 66, 67, 68, 69], while work on spatio-temporal reasoning has focused on the recognition of actions in videos [9, 10, 8, 45, 48, 72]. In the domain of e-learning, different works on the emotional recognition of learners, which could be used to determine whether the learner comprehends a lecture or not, have been discussed [214, 25, 12, 27, 215].

Due to the considerable successes reported in reasoning tasks by various studies using static images and video data with question-answering, the current study recognised that the potential to apply this form of reasoning to the educational domain, especially in e-learning, would be a big step forward. To successfully carry out the task of spatio-temporal reasoning in the domain of e-learning, this study intends to use video samples from learners in an online setup, CNN-BiLSTMs to estimate their facial emotion and affective states over a period of time, mapp the estimates to positive, negative and neutral, and generates narrative reports, in addition to Q&A and use these to reason about their comprehension about a lecture.

To the author's knowledge, no existing work in the e-learning domain has explored the use of reasoning using a combination of learners' estimated facial emotions

and affective states, mapping their estimates of emotion and affective states onto learning affects to generate narrative reports to be used for incorporating reasoning through Q&As. Q&A feedback through reasoning will assist teachers to adapt their lecturing style/content whereas this feedback will assist learners to find gaps in their comprehension of the material. The outcome of this feedback mechanism is expected to be a paradigm shift to improve online teaching and learning delivery systems, especially since e-learning is gaining popularity in the education arena.

Earlier work on reasoning that used video questions and answers focused mainly on reasoning about different actions, activity recognition, counting and other general human and animal behaviours. This study looks at diversifying the use of reasoning principles in the education sector by incorporating the idea of reasoning into an e-learning system. Learners' evaluation during teaching and learning is crucial to both teachers and learners and is therefore an area that needs much attention, especially in the e-learning domain, which lacks direct contact with learners. Some present e-learning systems concentrate on facial expression recognition and affective state recognition for evaluating learning affect. However, reasoning through narratives and question-and-answer phrases is not taken into account when estimating learners' learning affect. This study uses facial cues, such as emotional expressions and affective states from videos and extends it by incorporating reasoning through narratives and questions and answers feedback into existing e-learning systems. Such a pipeline that includes capturing learners' facial behaviour, processing and generating reasoning-based feedback has the potential to bring an online learning system closer to traditional classroom learning.

This study aims to achieve the above by designing a CNN-BiLSTM model using an existing video facial expression dataset for initial training and validation of the framework and a question-and-answer system module, using a pre-trained BERT-QA bot model. The proposed system of reasoning and its application to emotion analysis was inspired by the existing work in spatio-temporal reasoning and emotion, and affective states analysis on the e-learning platform.

To carry out the experiments for this dissertation and to test the proposed system for further verification of results reported, existing datasets were adopted for the training and testing experiments, while real-life data was acquired from learners. DISFA+ which is an extension of DISFA dataset [107] and DAiSEE which are spontaneous expressions datasets were selected to carry out the experiments. DISFA+ consists of labels of basic expressions of anger, disgust, fear, happy, sadness, surprise and neutral. DISFA+ dataset has been used for emotion-related experiments by researchers and has reported remarkable results [216, 215, 217]. Dataset for the Affective States in E-Environment (DAiSEE) datasets used for the affective states' estimation, consists of labels of affective states, which include, boredom, confusion, engagement and frustration. Authors in [218, 219, 220] have used the DAiSEE data set for their affective states-related experiments and have reported remarkable results. This study chose DISFA+ and DAiSEE for emotion estimate and affective state estimation because of the remarkable results provided through them.

CNN's hierarchical structures and various filter sizes afford these networks the ability to detect high, medium, and low-level features [221]. Using the concept of transfer learning, which means using a pre-trained CNN, will allow us to build

models in a time-saving manner [222]. This study adopted the concept of transfer learning for the feature extraction of the emotion and affective state samples using using a pre-trained CNN feature extractor: ResNet-50 and VGG-19. Because of the quality of BiLSTM, this study considers it to be an appropriate tool for our video sample classification for estimating learner's learning affects. BiLSTM is a type of LSTM structure that aids in learning long-term dependencies, which is suitable for sequential data and has the potential to increase the accuracy of a model [223, 224]. Relevant feedback is provided through narrations and questions and answers phrases using Bidirectional Encoder Representation Transformer (BERT) model. BERT is a language model that has produced cutting-edge results in a wide range of NLP tasks, including question-answering tasks, natural language inference, and others [225]. Details of these models are provided in chapter 3 of this dissertation.

2.8 Chapter Summary

The use of facial emotion and affective state recognition in e-learning has received a lot of attention, and various studies have sought to improve these existing systems. One of such is this proposed study, which seeks to employ spatio-temporal reasoning using facial emotion recognition, affective state recognition, and a NLP mechanisms to perform reasoning of learners' emotional states using Q&A phrases. To establish a relationship between the facial emotion, affective states discussed and the e-learning system, it was pertinent to discuss some research done in this area. A significant change in recent times has been the advent of machine learning and its integration into teaching and learning, which has resulted in the development of various models that can mimic the human teacher and have the potential to enhance e-learning platforms.

In this chapter, various literature that contributed to the study on reasoning tasks using static images, spatio-temporal reasoning using video data, their methods of approach, and their results have been examined. The chapter further discussed the study of behaviour in the education domain, and the gaps that necessitated the study were presented. Learning affect's overview and the importance of detecting the learning affects of learners were presented. The various approaches to affects analysis, affects categorizations/mappings, and the different datasets available for affects analysis have been discussed. Pre-processing methods, feature extraction techniques, and classification methods, which are the techniques used in the analysis of affects were also discussed. The background to this study is presented and discussed in chapter 3.

Chapter 3

Background

3.1 Introduction

The background techniques and architectures employed for this research study are explained in this chapter. Sections 3.2 discuss an overview of facial expression, 3.3 affective state, and 3.4 E-learning respectively. The video input of the datasets utilized in this study, which includes DISFA+ and DAiSEE datasets, is described in 3.5. The feature extraction using the transfer learning approach, which is the ResNet-50 and VGG-19 pretrained CNN networks, are discussed in 3.6. while section 2.5.3 details the classification strategies used, including the BiLSTM classifier and 2.6 details the feedback mechanism that was utilised, which includes the BERT-Q&A bot model architecture. Section 3.9 talks about the measurements used to gauge performance and the chapter summary is given in section 3.10.

3.2 Facial Expression

Emotions are spontaneous movements that occur over time and space when one or more facial muscles are activated [226]. Facial expressions are the most appropriate form of non-verbal communication and offer a lot of information [227]. Facial expressions have been shown to be universal, which means that the expression in one region is the same expression everywhere. Ekman set out to verify this further in [96], utilizing people from different cultures and regions to see if the emotional expression is the same for the same stimulus. He noticed that people reacted to stimuli in the same way, displaying the same facial expressions depending on their location. As a result, he came to the conclusion that facial emotional expression is universal. He further opined that there are seven basic emotions, which are happiness, sadness, fear, disgust, anger, contempt, and surprise [228]. Figure 3.1 illustrates these basic emotions.

Similarly, other studies in this field have revealed six basic emotions, which include happiness, sadness, anger, fear, disgust, and surprise [229, 230]. Some of these identified emotions can be attributed to the learning environment. Authors in [125] also proposed a real-time model for understanding learners' facial emotions and detected six different kinds of emotion categories, namely: sad, surprised, happy,

Figure 3.1: Six basic emotions.
Source: [228]

anger and disgust. These authors further explained that the learner's emotional data could be retrieved from a series of video frames. As a result, established that accuracy and speed must be balanced to perform real-time facial emotion analysis, especially in an e-learning environment.

3.3 Affective State

The experience of feeling the underlying emotional state is referred to as an affective state [231]. Affective states are longer-lasting mode states and are the result of a collection of experiences rather than a single stimulus [232]. Certain affective states during learning may enhance or hinder learner learning [233]. Based on observations, the human face reveals that certain emotions, such as anger, surprise, disgust, sadness, confusion, and joy, correspond to particular facial expressions. However, facial expressions can be very delicate, and some of them are unmistakably connected to affective states that develop during learning [17]. Kort et al. in [17] sought to identify those affective emotional states and axes that seem to be most relevant to the learning processes — both for voluntary learning (the intrepid learner) and for involuntary learning (the reluctant learner). They developed a slide to represent these affective emotional states as shown in Figure 3.2. On this slide, they presented several affective states as well as some complex phrases for different levels of intensity. For this study, boredom, engagement, frustration and confusion affective emotion states are used for the experiments.

3.4 E-Learning

E-learning means instructions that are disseminated electronically through a web browser, the internet, or any other medium. It is the organized, legitimate use of an electronic machine to support the learning process and the web-based integration of instructional practices and internet abilities to direct learners towards a specified level of knowledge in a specified domain [234]. Generally, E-learning is education that is received over the internet through smart devices like computers and other

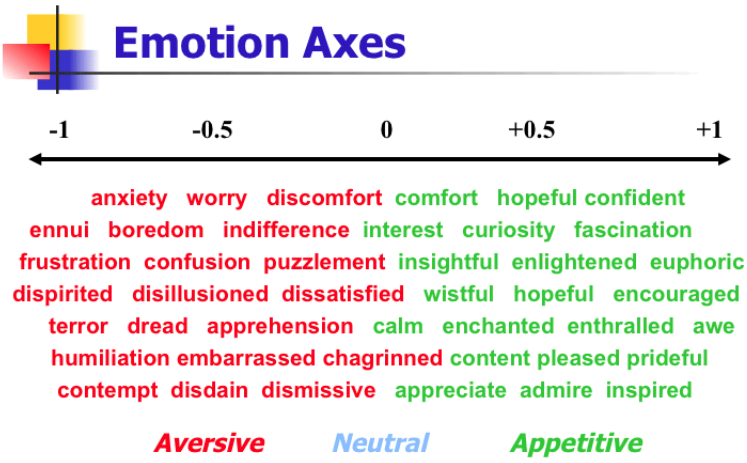


Figure 3.2: Affective emotion axis [17]

related devices. Previously, some related studies in the area of emotion expression and with respect to e-learning have been discussed [234]. Therefore, it will be important to elaborate more on the term "e-learning". The authors of [235] explained that it is pertinent that there be an evolution in the methods of approach to the learning process. The internet had been a tool of choice for the realization of this evolution because internet technologies have been of higher patronage in recent years. These technologies, which in turn have brought about the stability of e-learning, support many practices such as learning, collaboration, communication and interaction, among others. E-learning has become an indispensable word in the educational technology domain. At the advent of e-learning, it was called online "learning", and "virtual learning", but the technology has remained the same since then [236].

Among all these modalities examined, we selected facial cues for the research in this study. Facial cues were selected because they are the most appropriate source of nonverbal communication, and nonverbal communication accounts for up to 93% of the impact of any verbal message [237, 238]. As a result, facial expression and affective state detection have become active topics of research for e-learning environments, which has triggered the use of facial expression and affective state recognition for study in this dissertation.

3.5 Datasets

Spatio-temporal reasoning for estimating students' learning affect is a sequence classification problem. Sequence classification is used to classify a predictive modelling problem with a sequence of inputs over space and time. The task is to predict a category for the sequence [239], which is video emotion and affective states classification. Thus, the proposed work inhere used video datasets as input data for the Facial Emotion Recognition (FER) experiments. A fundamental part of affect-aware

systems is automatic FER. Due to a lack of annotated spontaneous data, the majority of earlier automated FER systems were trained on posed facial expressions. On the other hand, we deal with actual facial expressions in real-world applications. This study, therefore, adopts DISFA+ and DAiSEE datasets, which were previously released and well-accepted facial datasets.

3.5.1 DISFA+ dataset

DISFA+ datasets contain labels of basic emotions of anger, disgust, fear, happy, surprise, sadness and neutral. DISFA+ dataset was requested for use for emotion estimation in this dissertation. Figure 3.3 presents some samples from DISFA+ dataset.

Figure 3.3: Samples from DISFA+ Dataset [18]

3.5.2 DAiSEE dataset

DAiSEE was the dataset utilized for affective state estimation. It consists of labels of affective states of boredom, frustration, engagement and confusion. DAiSEE is a freely available dataset. Figure 3.4 shows some samples of DAiSEE dataset.

In light of the literature discussed, this makes the DISFA+ and DAiSEE datasets appropriate sources for the proposed objectives of this study. DISFA+ and DAiSEE datasets have been detailed in 2.4.5.

3.6 Feature Extraction Method used in this Work

One of the major steps towards effective emotion/affective state recognition is to represent the facial data being used effectively. Representing an image in machine learning is referred to as feature extraction [240]. CNN has been used by various researchers for feature extraction over the past decades, and this has sparked a lot of interest, both for image classification and feature representation [241, 149]. CNN networks have been trained using different datasets such as ImageNet, MNIST, CIFAR-10, among others, to extract features from various images that are used for

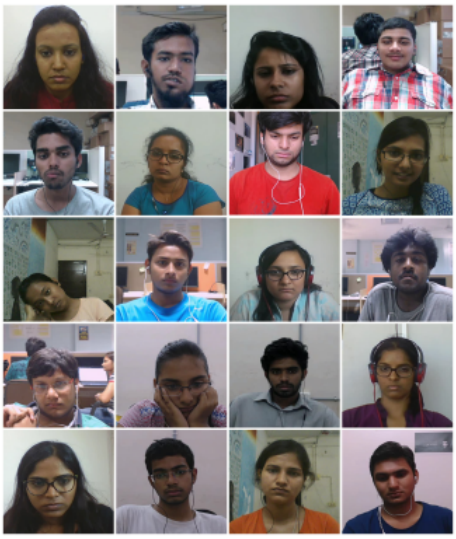


Figure 3.4: Sample video frames from DAiSEE [19]

classification purposes[135]. Further studies have brought about an improvement on this conventional feature extraction method. The CNN is pretrained on some datasets and subsequently, used to learn patterns from other datasets that it was not originally trained on. This method is known as transfer learning.

For the experiments carried out in this study, we propose to use CNN pretrained networks for feature extraction of the sequence data samples of the videos, and a BiLSTM network approach for emotion and affective state classifications of the extracted feature samples. The CNN pretrained networks that were proposed for use for feature extraction of the sequence data samples are the ResNet-50 and VGG-19 that were pretrained on ImageNet.

3.6.1 ResNet-50 pretrained CNN

Deep CNN has produced a number of improvements in image identification and classification over the years. It is becoming conventional to delve deeper to solve more difficult problems and increase classification or recognition accuracy. However, due to issues such as the vanishing gradient problem, training deeper neural networks has been challenging [242]. The known vanishing gradient is one of the difficulties that Residual Networks (ResNet) handle. This result on the weight never modified its values, and no learning was done as a result because the gradients from which the loss function is derived readily drop to zero after repeated applications of the chain rule when the network is too deep. Gradients from later layers to the initial filters can flow directly through the skip connections with the residual layer. As a result, ResNet is employed to address the vanishing/exploding gradient issue. The technique this network uses is called skip connections [243]. The skip connection skips training from a few layers and connect directly to the output as illustrated in Figure 3.5.

ResNet-50 is one of the CNN versions of ResNet [20], which has 50 layers. ResNet

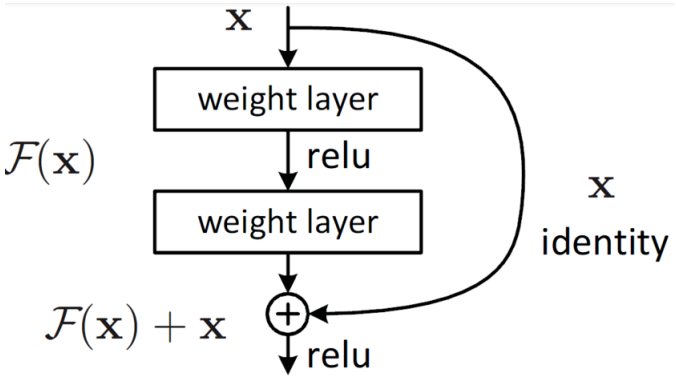


Figure 3.5: Residual Network [20]

was trained on the ImageNet database and could classify 1,000- categories of objects [244]. ResNet-50 consists of 48 Convolution layers, 1 MaxPool layer, and 1 Average Pool layer. Despite having 50 layers, ResNet-50 contains approximately 23 million trainable parameters, which is significantly less than other architectures. Through various literature surveys, ResNet-50 has been proven to perform more efficiently in the extraction of features as observed and reported by some authors and discussed in 2.5.2.

3.6.2 VGG-19 pretrained CNN

The Visual Geometry Group (VGG) at Oxford developed a pretrained CNN model which is a successor of the AlexNet. The VGG has several variants, of which VGG-19 and VGG-16 are parts. VGG-19 has 19 layers consisting of 16 convolution layers, 3 fully connected layers, 5 Maxpool layers, and 1 SoftMax layer. Figure 3.6 illustrates the architecture. It also consists of 19.6 trainable parameters, pretrained on over a million images from the ImageNet database [21]. The architecture used a fixed-size (224×224) RGB image as the network's input, therefore the matrix had that shape (224,224,3). The main preprocessing carried out on VGG-19 involved calculating the mean RGB value over the whole training set and subtracting it from each pixel. The use of kernels of (3×3) size and a stride size of 1 pixel allowed them to completely cover the image's notion. To maintain the image's spatial resolution, spatial padding was applied. Stride 2 was used to conduct max pooling over a 2×2 pixel window. Rectified linear unit (ReLu) was used to introduce non-linearity and speed up computation after this to increase the model's ability to categorize data. This study fine-tunes the VGG-19 for the extraction of the affective state's samples feature using transferred learning.

3.7 Classification Method used in this Work

For the classification of extracted features, a bidirectional LSTM (BiLSTM) was used and is discussed as follows.

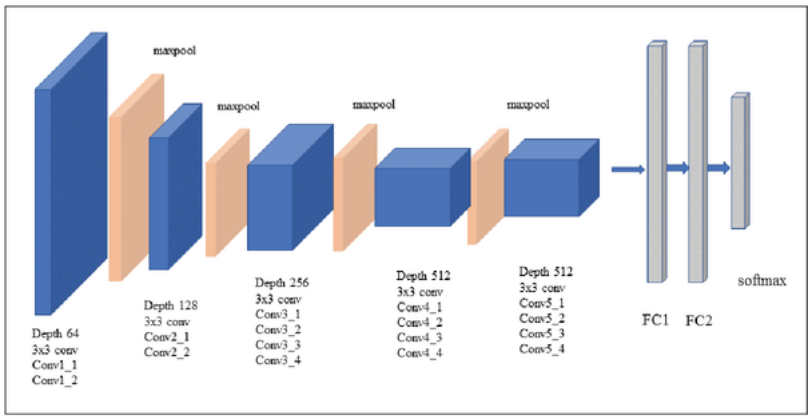


Fig. 3. VGG-19 network architecture

Figure 3.6: VGG-19 Architecture [21]

3.7.1 Architecture of LSTM

LSTM is a type of RNN designed for sequence problems [22]. With the adoption of RNN to solve sequence problems, the network only remembers data content for a short while, and when a high volume of data is fed into the network, the information gets lost. This problem was tackled with this new version of the RNN called LSTM [22, 245]. LSTM was first introduced by Hochreiter and Schmidhuber in 1997, and since then, many researchers have improved on it, and it is widely used [64]. For relation networks to add new information to a network, it changes the existing information completely by implementing a function, thereby modifying the whole information. This implies that no consideration is being made for 'important' and 'not so important information. However, in the case of LSTM networks, it would make a change to the new information by multiplication and addition actions. With LSTM, information flows through a mechanism called a cell state, and this is why LSTM can selectively remember or forget information.

The LSTM is made up of memory blocks called cells [22] where two states are transferred to the next cell; the cell state and the hidden state. The memory blocks remember information, and manipulations of this memory are done through 3 primary mechanisms, called gates. These are; the forget gate, the input gate and the output gate as Figure 3.7.

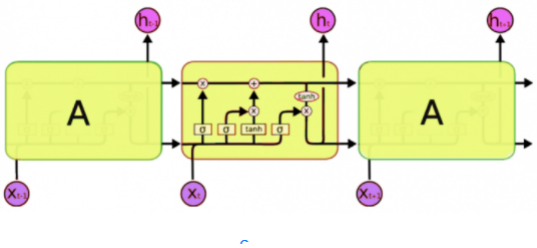


Figure 3.7: The Architecture of LSTM [22]

The Forget gate: This gate is responsible for replacing or forgetting a subject in a previous sentence or phrase with the new subject in the new sentence. The forget gate removes information from the cell state that is, the information that LSTM

does not need to understand things. For example, assume that an LSTM is being fed with the sentence: "Jerry is a good man. Fabian, on the contrary, is a bad guy". Immediately the full stop after "man" is encountered in the first sentence, and the forget gate observes a change of context in the preceding sentence. Apparently, the subject of the first sentence, which is "Jerry" is forgotten; the position of the subject is emptied. When the next sentence is encountered, the position of the subject is given to the new subject "Fabian". This process of forgetting the subject in the first sentence whenever a new sentence is encountered is carried out by the forget gate. The forget gate is responsible for removing information from the cell state [22]. This implies that information the LSTM does not need to understand is removed by the forget gate.

The Input gate: The task of adding new information to the cell state is done through the input gate. Only relevant information in a sentence is added to the cell state through the input gate.

The Output gate: Some information that is passed along the cell state might not be fit for being output. The output gate is responsible for selecting useful information from the current cell state and sending it out. Authors in [64, 246] proposed some LSTM equations with the forget gate given below:

Lowercase variables in this equation represent vectors. The matrices W_q and U_q contain the weights of the input and recurrent connections, respectively, and the subscript q can be either the input gate I output gate o, forget gate f, or memory cell$c$, depending on the activation being calculated.

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \tag{3.1}$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{3.2}$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{3.3}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \tag{3.4}$$

$$h_t = o_t \circ \sigma_h(c_t) \tag{3.5}$$

Here, the initial values are $c_o = 0$ and $h_o = 0$ and the operator $\circ$ denotes the Hadamard product (element-wise product). The subscript t indexes the time step [64, 246]. The variables are given as :

$x_t \in R^d$: Input vector to the LSTM Unit

$f_t \in R^h$: Forget gate's activation vector

$i_t \in R^h$: Input and update gate's activation vector

$o_t \in R^h$: Output gate activation vector

$h_t \in R^h$: Hidden state vectors, also known as the output vector of the LSTM unit.

$c_t \in R^h$: cell state vector

$w \in R^{hxd}, U \in R^{hxh}, and b \in R^h$: Weight matrices and bias vector parameters which need to be learned during training, the superscripts d and h are the number of input features and the number of hidden units, respectively.

σ_g: Sigmoid Function

σ_c : Hyperbolic tangent function. [64]

3.7.2 Architecture of BiLSTM

Bi-directional LSTM is an advanced form of the LSTM network, consisting of two LSTMs. The first LSTM is applied to the data from the input sequence in the forward direction, while the second LSTM takes the reverse structure of the input sequence. This type of LSTM structure aids in learning long-term dependencies and has the potential to increase the accuracy of a model as described by the authors in [223, 224]. The hidden forward layer h_t^f processes the input in ascending order, that is, t = 1,2,3....T, while the backward hidden layer h_t^b processes the input sequence in descending order, that is, t = T,.....3,2,1. Lastly, both layers are combined to generate the output Y_t ([182]). A Bi-LSTM is implemented using the following equations:

$$h_t^f = \tanh(W_{xh}^f xt + W_{hh}^f h_{t-1}^f + b_h^f) \tag{3.6}$$

$$h_t^b = tanh(W_{xh}^b xt + W_{hh}^b h_{t+1}^b + b_h^b) \tag{3.7}$$

$$Y_t = W_{hy}^f h_t^f + W_{hy}^b h_t^b + by \tag{3.8}$$

W is a weight matrix, while tanh is a hidden layer activation function and W_{xh} is a weight connecting input (x) to the hidden layer (h), and b is a bias vector for both the forward and backward hidden layers h_t^f and h_t^b as shown in equations (3.6) and (3.7). Figure 3.8 shows the architecture of BiLSTM.

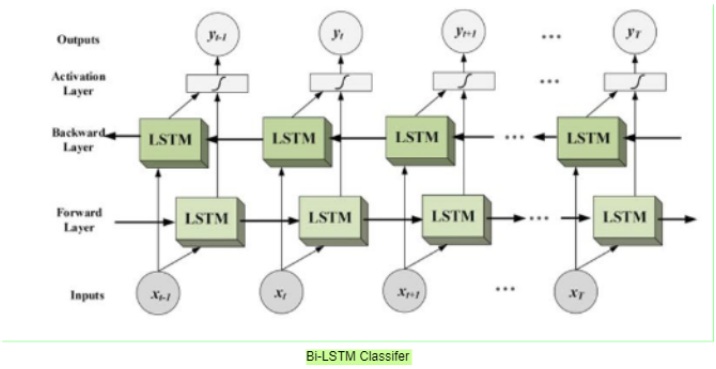


Figure 3.8: The Architecture of BiLSTM [23]

3.8 Feedback Mechanism used in this Work: BERT model

To provide feedback through questions and answers, showing reasoning on learning experienced during teaching and learning based on estimated facial cues, estimation of learning affect, and generated narrative reports, this study adopted the BERT-Q&A-bot. BERT is a language model that has produced state-of-the-art results in a wide variety of NLP tasks, including question-answering tasks, natural language

inference, and more [225]. BERT is a language representation model that is designed to extract pre-trained deep bidirectional representations [24]. BERT uses a transformer. A transformer is an attention mechanism that learns the contextual relationships between words in a text. Transformers use two separate mechanisms namely, an encoder that reads the text input, and a decoder that produces a prediction for the task. Only the encoder mechanism is required because BERT's aim is to produce a language model. The transformer encoder reads the entire sequence of words at once. This is bidirectional, and this characteristic allows the model to learn the context of words based on all of its surroundings [225].

The BERT framework comprises two stages: pre-training and fine-tuning [24]. Throughout the pre-training phase, the model is trained on unlabeled data using a variety of pre-training tasks. The BERT model is initially initialized with the pre-trained parameters, and then all of the parameters are fine-tuned using labelled data from the downstream tasks. Each downstream task has its own set of fine-tuned parameters. An example is the Question Answering BERT that is known to have a unified architecture across different tasks, hence its distinctive features.

BERT Architecture

BERT's model architecture is a multi-layer bidirectional transformer encoder as described in [247]. In the model architecture, L is the number of layers, H is the number of hidden layers, and A is the number of self-attention heads. Authors in [24] reported that BERT has two sizes, these are:

$BERT_{Base} \quad with \quad L = 12, H = 768, A = 12, Total \quad Parameters = 110M$

and

$BERT_{Large} \quad with \quad L = 24, H = 1024, A = 16, Total \quad Parameters = 340M$

BERT model was trained on 2.5 million words of data from Wikipedia and fine-tuned on small data NLP tasks like question-answering. According to the authors of [248], for question-answering tasks, an already trained model can be used and produce reasonable results even when the text is from a completely different domain. To achieve a good result, the BERT model, which has been fine-tuned on the SQuAD can be used. Stanford Question Answering Dataset v1.1 benchmark (SQuAD) is a reading comprehension dataset consisting of questions posed by crowd workers of a set of Wikipedia articles, where the answer to every question is a segment of text or span from the corresponding reading passage, or the question might be unanswerable.

Pre-Training BERT

Before discussing the pre-training methodology, there is a need to have a quick glance through the input/output representations used in BERT. For the input and output representations, BERT is able to represent both a single sentence and a pair of sentences. It can represent sentences such as a Q & A, in a single token sequence to handle any task. A token is referred to as an instance of a sequence of characters that are put together as a relevant semantic unit for processing in a certain document [249]. In contrast to a real language sentence, a sentence is a length of continuous

text. [24]. A sequence, therefore, is the input token sequence to BERT which is either a sentence or two sentences packed together. The 30,000 token vocabularies were used with WordPiece embeddings. WordPiece embeddings use a tokenization technique that alternates between character-level encoding and full-word encoding to boost coverage for the majority of vocabulary words while significantly reducing the vocabulary size. The first token of each sequence is always a special classification token ([CLS]). For classification tasks, the final hidden state corresponding to this token is used as the aggregate sequence representation. A single sequence has several combinations of sentences. There were two ways to differentiate the sentences. They were first divided using a unique token ([SEP]). Second, each token was labelled with a learnt embedding that designates whether it belongs to sentence A or sentence B as shown in Figure 3.9

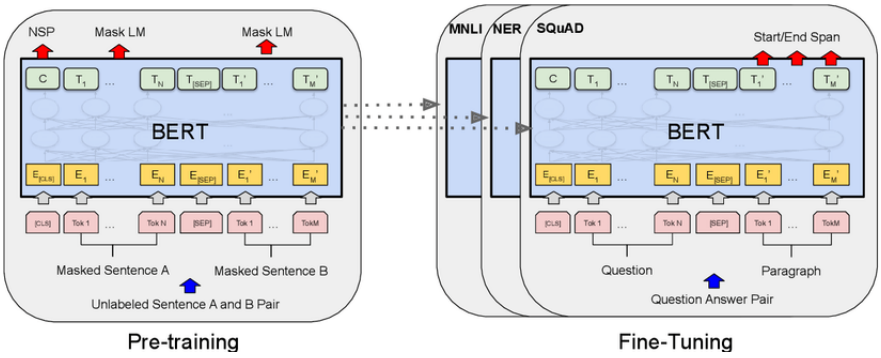


Figure 3.9: BERT Training and Fine-tuning [24]

The input embedding is denoted as E, the final hidden vector of the special classification token [CLS], is denoted as $C \in R^H$, and the final hidden vector for the i^{th} input token as $T_i \in R^H$. To create the input representation for a given token, the corresponding token, segment, and position embeddings are added together. Figure 3.10 shows how this construction is visualized.

In Figure 3.10, [CLS] is known as the 'classification' token and is used at the beginning of a sequence. [SEP] is a token that indicates the separation of 2 sequences. [MASK] is used to indicate a masked token in the MLM task.

Segment embeddings: Segment embeddings are used to determine the sequence to which the token belongs when there are many sequences in the input that are divided by the token [SEP]. To determine the token's sequence, these are combined with the model's initial Word embeddings. Using WordPiece embeddings, the tokens provided to the BERT model are tokenized.

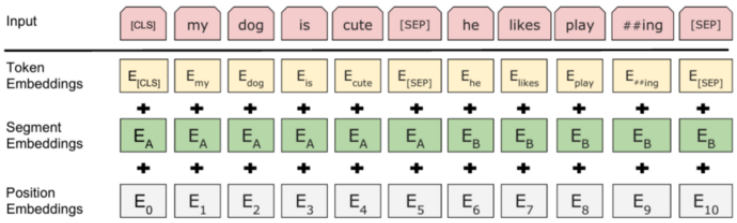


Figure 3.10: BERT input representation [24]

BERT was pretrained using two unsupervised tasks. These include Masked Language Model (MLM) and Next Sentence Prediction (NSP) tasks.
MLM means the training of a bidirectional representation in this task, 15% of each sequence's tokens are randomly masked (replaced with the token [MASK]). The other tokens in the sequence are used to train the model to predict these masked tokens [24].
NSP: Understanding the relationship between two sentences, which is not specifically represented by language modelling, is the foundation for several crucial downstream tasks, including Q&A and Natural Language Inference (NLI). To train a model that understands sentence relationships, a binarized next-sentence prediction task that can be generated from any monolingual corpus is pretrained. Two input sequences are given (separated using the [SEP] token, and Segment Embeddings are used). It is a binary classification task that predicts if the second sentence succeeds the first sentence in the corpus. For each pre-training example, the sentences A and B are randomly selected, with 50% of the time B being the sentence that really follows A (labeled [IsNext]) and 50% being a random sentence from the corpus (labeled [NotNext]). In Figure 3.9, C is used for the next sentence prediction.

Fine-Tuning BERT

Modeling any downstream tasks by BERT was made possible by the self-attention mechanism in the transformer. In order to unify two phrases, BERT uses a self-attention method. A concatenated text pair with self-attention effectively contains bidirectional cross-attention between two sentences. Each task's input and output are connected to BERT, and all of the parameters are adjusted from beginning to end. At the input end, sentences A and B from pretraining are equivalent to (1) sentences pairs in paraphrasing; (2) hypothesis-premise pairs in entailment; (3) a degenerate text pair in text classification or sequence tagging, and (4) question-passage pairs in Q & A [24]. At the output end, the token representation is fed into an output layer for token-level tasks such as the question-answering task. The Stanford Question Answering Dataset (SQuAD) is a collection of 100 thousand crowd-sourced Q & A pairs [250]. The aim is to predict the answer text span in a paragraph from Wikipedia when given a question and the passage that contains the answer. Passages are described as articles, paragraphs, and sentences by the authors in [251].

In the Q & A and answering task, questions and passage is input as a single package sequence, where 'question' is the A embedding, and the 'passage' is the B embedding. A start vector $S \in R^H$ and an end vector $E \in R^H$ are introduced during fine-tuning as shown in Figure 3.9. Dot products between Ti and S are used to calculate the probability that words i will begin the answer span, which is then multiplied by a softmax across each word in the paragraph:

$$P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}} \tag{3.9}$$

To compute the score of the answer span and the maximum scoring span from position i to position j is given as $S \cdot T_i + E \cdot T_j$, x given that $j \geq i$ is used as a prediction. The sum of the log-likelihoods of the correct start and end positions is

the training objective.

To achieve the question-answering task in this proposed study, transfer learning concept was used in the BERT model. The transfer learning approach was achieved by replacing the input layer of the fine-tuned model with the question and passage designed for this study. $BERT_Large$ model that was fine-tuned on SQuAD v2.0 was used. The model aimed to predict the answer to the question based on narrative feedback. BERT is pretrained to accept 512 tokens as the maximum number of tokens in the form of text. This implies that both the question and the passage are a maximum of 512 tokens.

3.9 Metrics used for Performance Measurement

Evaluation is an important aspect of any research work. The most crucial question that arises once a model has been trained is how good the model is. This translates into the most crucial work in machine learning, i.e., evaluating a model to determine the accuracy of the predictions. In this study, two levels of evaluation were considered. The first level considered the correctly estimated facial emotion, affective states and learning affect, and the second level considered the correct estimation of facial emotion, affective states and learning affect through live testing and questionnaire. The evaluation metrics used for the first level of emotion and affective states experiments in this study are the confusion matrix, accuracy, precision, recall, and the F-1 score. These are defined as follows:

- **Confusion Matrix**: A confusion matrix is a table used to describe the performance of a classifier. It comprises rows and columns in which the row represents the actual class of emotion, while the column represents the predicted values. All the correctly predicted values are shown in the diagonal of the table. Prediction errors can be easily located because they will be values outside the diagonal [252]. The confusion matrix is calculated as follows:

$$\frac{M_{ii}}{\sum jM_{ji}} \tag{3.10}$$

- **Accuracy**: The most intuitive performance metric is accuracy, which is defined as the ratio of correctly predicted observations to all observations [252]. Accuracy is calculated as:

$$Accuracy(A) = \frac{TP + TN}{TP + FP + FN + TN} \tag{3.11}$$

- **Precision**: The ratio of accurately predicted positive observations to total expected positive observations are known as precision[252] and it is calculated as:

$$Precision(P) = \frac{TP}{TP + FP} \tag{3.12}$$

- **Recall**: Recall is the fraction of cases where the algorithm correctly predicted

i out of all cases, which are labelled as i [252]. Recall is calculated as:

$$Recall(R) = \frac{TP}{TP + FN} \tag{3.13}$$

- **F1- Score** is the weighted average of precision and recall [252]. F1- Score is calculated as:

$$F - 1 \quad Score = \frac{2 * P * R}{P + R} \tag{3.14}$$

From the calculations, TN,FN,TP and FP can be established as: the correctly predicted positive values are denoted by TP, and the correctly predicted negative values are denoted by TN. When the actual class is "no" and the expected class is "yes" the outcome is FP; when the actual class is "yes" but the projected class is "no" the result is FN [252].

The second level of the evaluation was based on the live test results and output statements from students during the live test, which was done through a survey using questionnaires. The students were asked questions and each question tries to evaluate the proposed system. The different responses were analysed and used to evaluate the proposed system. The detailed analysis is presented in chapter 5.

3.10 Chapter Summary

This study considered facial cues such as emotions and affective states as providing some basic information that a teacher needs to analyse the quality of his or her teaching during e-learning. Getting appropriate deep learning mechanisms including feature extraction methods with appropriate classification algorithms, to estimate these cues is of crucial importance.

The background information that establishes the context for this study is presented in this chapter. Information such as the basic facial expressions provided by researchers was discussed. Facial expressions including anger, disgust, fear, happiness, surprise, sadness, and neutral, were highlighted. Affective states of boredom, frustration, engagement, and confusion were presented based on literature, and E-learning system was also discussed.

The datasets used for emotion and affective state experiments, which included the DISFA+ and DAiSEE datasets, were highlighted in this chapter. A transfer learning approach to feature extraction was presented, which included pretrained CNNs such as ResNet-50 and VGG-19. The existing classification methods used for the emotion and affective state estimators, which include BiLSTMs, were also discussed. The BERT model fine-tuned for the question-and-answer output aspect of this work was also presented. The metrics that were used for the measurements of the performance of the models, which include the confusion matrix, accuracy, precision, recall and F1 score were also presented.

This study expects that the datasets, feature extraction method, and classification algorithm used will be suitable for training the model to provide the required output when tested in real-world scenarios. Similarly, report generation and the

inclusion of the BERT-Q&A bot model are expected to efficiently handle the feedback aspects of this study. Chapter 4 discusses the methodology used to achieve the study's objectives.

Chapter 4

Research Methodology

4.1 Introduction

This chapter outlines the techniques used to establish a system that estimates the learning affect of a student and incorporates questions and answers to provide feedback to teachers and students. The lack of reasoning in the e-learning system that served as motivation for this study is discussed in section 4.2. Section 4.3 describes the proposed framework. Components of the framework that include the teacher's input, the student's input, CNN-BiLSTM to extract features and classify emotions and affective state samples were described. The section also described the learning affect estimation, and report generation modules of the framework. A BERT-Q&A bot module that was used to accept predefined questions and generates answers was described. The data and its format for the experiment set-up are presented in section 4.4, while section 4.5 covered the discussion of emotion and affective state classifiers, feature extraction techniques, training and validation experiments, and results. Section 4.6 discussed the initial mapping based on the literature survey. These mappings include Emotions to Learning Affect (ELA) mapping and Affective state to Learning Affect (ALA) mapping. The feedback mechanism, which includes the narrative report generation and Q&A using the BERT-Q&A bot, was discussed in section 4.7, while the live test experiment conducted on students using the proposed model and the ethical considerations were discussed in section 4.8. The chapter summary is found in section 4.9.

4.2 Problem Analysis

Including reasoning into the e-learning system to improve the estimation of student comprehension, and be used as a means of feedback generation to teachers and students using their facial emotions and affective state analysis, remains an open problem. These problems are examined in this section. Previously, reasoning with question-and-answer tasks had been explored using artificial intelligence but their application to the education domain especially to e-learning was not fully explored. In the e-learning setup, studies have used video facial cues datasets for emotion and affective state recognition, but these studies lack NLP-based feedback mechanisms.

Therefore, this study proposes to extend the current setup in the e-learning domain. This study aims to achieve this by incorporating facial emotion and affective state recognition, learning affect estimation, reasoning through narrative feedback generation, and Q&A, where students/teachers can pose questions. The set of questions will focus primarily on the student's comprehension of the material. The goal is to generate reasoning-based answers as output while considering the context of the estimated emotion, affective state, learning affect, and metadata from the video lecture (i.e., labels, length, etc.).

4.3 Proposed Model Framework

The framework for the proposed study is presented here. Video recordings depicting a user's facial expressions will form input to the framework. (either saved or live) will be inputted into the framework. Input video streams will be fed into pre-trained CNN, which will extract the discriminative features. The extracted feature from the sequence of frames will be used as input for the emotion classifier BiLSTM and the affective state classifier BiLSTM. The estimates of emotion and affective state will be further mapped to positive, negative, and neutral learning affect as per existing literature. The mappings were further validated using cross-testing of trained classifiers and existing mappings. As part of the feedback mechanisms, narrative text reports were generated using templates, estimates of emotions, affective states, learning affect, and metadata from the video lecture. Narrative feedback was generated for both students and teachers. To enhance the usability of the framework, we also proposed the inclusion of a Q&A module. To handle questions which are text-based, a BERT-Q&A bot model was integrated. A predefined set of questions was formulated, on behalf of students and teachers. A user (teacher/student) then selects a specific question to which an answer would be generated by the BERT-Q&A bot model. The goal was to embed reasoning about the learning affect and overall comprehension of a student. Figure 4.1 shows the proposed model framework for spatio-temporal reasoning for estimating the student learning affect. This proposed model is presumed to improve the current e-learning platforms. The proposed framework's components are described in 4.3.1.

4.3.1 Framework description

The proposed spatio-temporal reasoning framework consists of two parts: one is the framework for a single student's input and feedback, while the other is for a group of students through which the teacher can have access to his or her feedback.

(i) Teacher: A teacher posts his lecture online for students during e-learning.

(ii) Students/Learners: The students/learners are the subjects who watch the online video lecture, while their facial emotions/affective states are captured by the system. The facial input was then split into a sequence of frames.

(iii) Emotion/ affective state feature extractors: In these modules, pre-trained ResNet-50 extracts emotion features from the frame sequence, while the pre-trained VGG-19 CNN extracts features of affective states from the frame sequence.

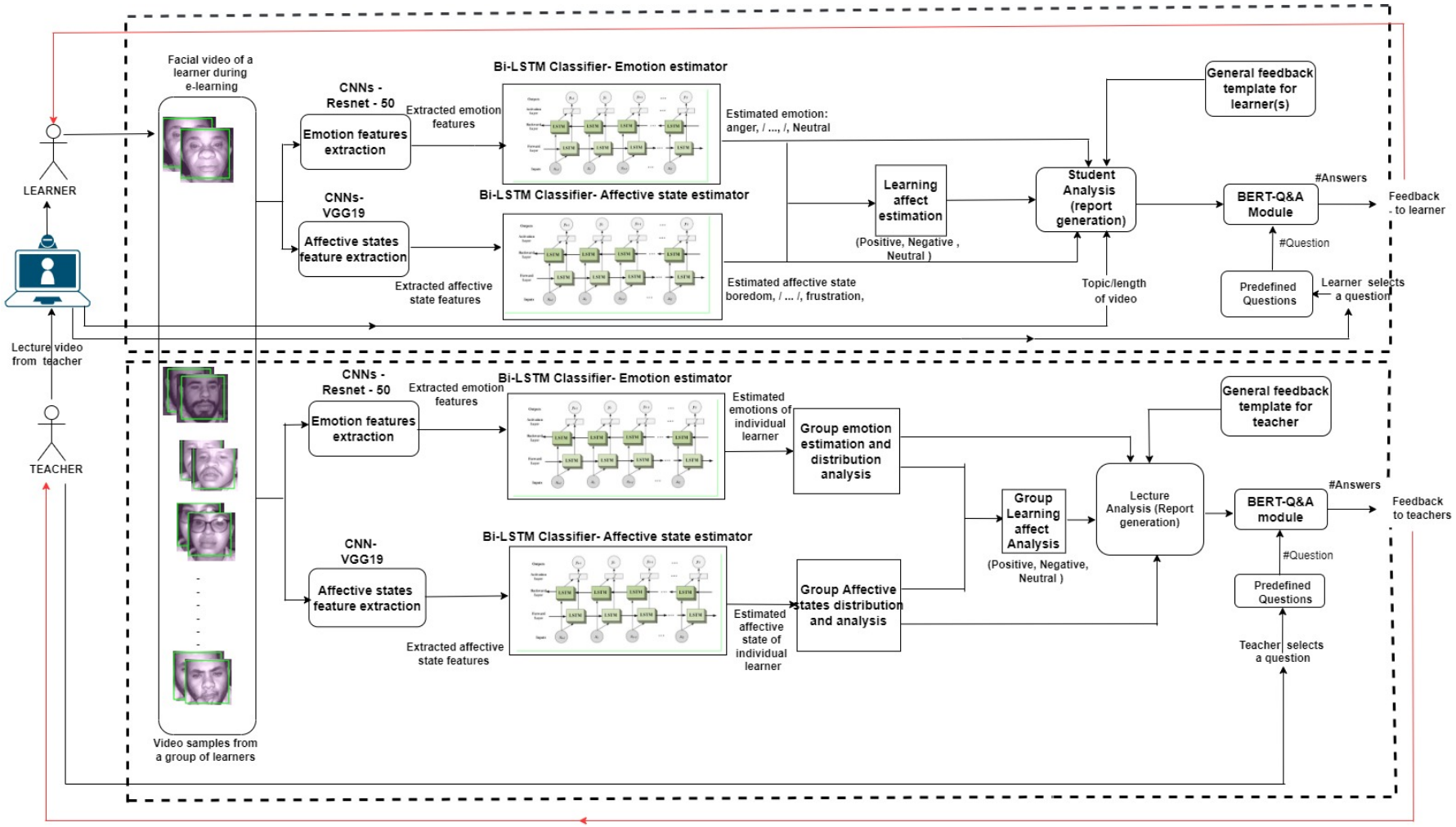


Figure 4.1: Proposed Spatio-Temporal Reasoning framework

This applies to both an individual student and a group of students as shown in the framework.

(iv) Emotion estimator (using BiLSTM Classifier): This module estimates the basic emotional states of the students during learning using the emotion features extracted by the Resnet-50 feature extractor as input. This module is applied to both the single and group of students' frameworks.

(v) Affective state estimator (using BiLSTM Classifier): This module, which is applicable in both single and group frameworks, estimates the affective state shown during learning using the affective state feature extracted by VGG-19 as input.

(vi) Learning affect estimation: This module maps the various estimated emotions and affective states to learning affects, including positive, negative, and neutral, as verified in this study. A statistical module is utilized to produce aggregates of the emotions and affective states that students have experienced on an individual and group level.

(vii) Report generation: The report generation module generates a report from the output of emotion and affective state estimators and the learning affect module, along with the video topic and length, using templates. There are two templates in this case, one for an individual student's report and the other for the report of a group of students.

(viii) BERT-Q&A module: In this module, predefined questions from the teacher or students are captured and processed. Reasoning answers are generated based on the question asked, using input from the report generated from the students or the lecture's analysis as a passage on which the context of the answers provided is based. The answers provided serve as feedback to the student or the teacher.

This study carried out the baseline experimental work on an Intel Core i7, 8^{th} generation processor with 16 GB of RAM using the Linux operating system. The algorithms were executed in a tensor-flow environment, and the training epoch and loss were visualized using a tensor board. The experiments' procedures are described in subsequent sections.

4.4 Data

The data and its format for the experimental set-up proposed in this study are described in this section. Video samples for baseline experiments were acquired from DISFA+ [18] and DAiSEE [19]. The study used a total of 26,603 sequence frames from DISFA+ video sequence, which were split for training, validation and test experiments. For the second experiment, a total of 39,215 sequence frames from the DAiSEE video sequences were acquired for training, validation, and testing of the affective state classifier. To validate the framework, live tests were performed that included a sample of 51 students from the University of the Witwatersrand. Appropriate steps were taken to acquire the ethics clearance for the study. This study summarizes the distribution of samples of datasets used for both classifications in Table 4.1.

Table 4.1: Distribution of data samples for the experiments

Dataset	Training	Validation	Test	Total Sample Size
DISFA+	18,247	6,083	2,274	26,603
DAiSEE	20,946	13,934	4,305	39,185

4.5 Classification Experiments

The following subsections provide in detail experimental setup for the task on hand. Section 4.5.1 describes the training and validation of the emotion classifier using emotion samples from DISFA+ followed by subsection 4.5.2, which describes the training and validation of the affective state classifier using samples from DAiSEE dataset.

4.5.1 Emotion Classifier

For the emotion classification experiment, the following steps were involved: (i) data pre-processing and feature extraction using pre-trained ResNet-50; (ii) classification of extracted features using BiLSTM classifier; (iii) Optimization using validation dataset. The framework for emotion classification is shown in Figure 4.2.

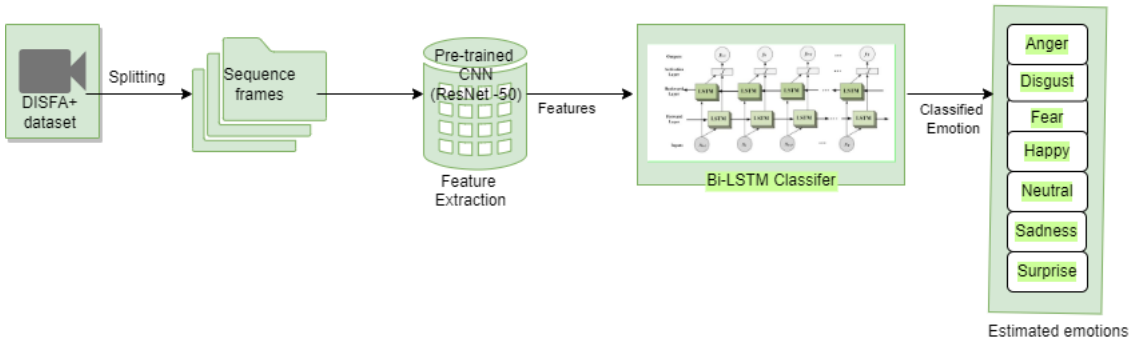


Figure 4.2: Emotion classifier framework

(i) Data pre-processing and feature extraction: DISFA+

The DISFA+ video data, which comprises sequences of labelled images of anger, disgust, fear, happy, sadness, neutral and surprise emotions, with 1280×720 pixels, was split into frames of sequence images. The sequence images were stored in HDF5 file formats, ensuring that the images were resized to $200 \times 200 \times 3$ pixels. The images were further pre-processed by expanding the dimensions and subtracting the mean intensity of the pixel RGB from the imageNet dataset (the original dataset on which the network was trained). In this study, a total of 26,603 sequence frames from DISFA+ video sequence was used. 24,329 sequence frames were used for training and validation with a split ratio of 75:25, while 2,274 sequence frames that were not part of training and validation were used to test the model. The discriminative

features were extracted using ResNet-50 pre-trained CNN. According to Algarni et al.[253], features were extracted for two reasons: first, more relevant information can still be recovered, contributing to accurate findings; and second, to minimize the dimensions of the data, better prepare the data for classification, and thus increase classification accuracy.

The CNN, which comprises 48 convolution layers, one maxPool and one average pool layer, was first loaded with the input images. The CNN accepts image input from the HDF5 file where the data was stored. The pre-trained ResNet-50 accepted input video sequences that were split into different frames of images. The frames reshaped to $200 \times 200 \times 3$ each, which is the input size required by the ResNet-50 networks, were added to a batch and passed through the ResNet-50 network for extracting discriminate features from the input images. The images forward propagated through the network until they reached the last max-pool layer, where the output shape's value of 2,048 was used as the feature vector for each sample. Therefore, the network produced a 2D vector of size 24,329 x 2,048. The extracted frame feature was passed as input to the BiLSTM network, which performed a facial expression classification.

(ii) Training and validation of emotion classification model

Understanding human facial expression plays a significant role in understanding emotions [254]. In this study, we adopted a BiLSTM network for emotion analysis. The extracted discriminative features were reshaped to a 3D array comprising the shape, sample size, sequence length, and input dimension. The input to the BiLSTM classifier is a sequence of frames, which is a $24{,}329 \times 2{,}048$ number of features from each frame. This was divided into training and validation datasets (18,247 and 6,083 samples) and then passed to the classifier using a batch size of 143 and a maximum epoch set to 50. Binary cross-entropy was used to identify the loss function, while the sigmoid function was used to compute the probability of the output labels.

(iii) Training and validation result: emotion classifcation

The BiLSTM network classifies the input feature, considering both the temporal and spatial aspects of the sequence samples, and validates the classification using validation datasets of 6,083 samples. An accuracy of 94.48% was reported on the validation dataset. The training and validation accuracy graph and the training and validation loss graph for the validation experiment results are shown in Figures 4.3 and 4.4, respectively.

The graph in Figure 4.3 shows the progress of the training as the number of epochs increases. During training, the accuracy reached its maximum at the 46th epoch, after which there was no improvement. Further training would cause an over-fitting of the model, and therefore, the model was stopped from further training at this point and the validation accuracy was reported. The emotions classified were anger, disgust, fear, happiness, sadness, surprise, and neutral. The graph in Figure 4.4 showed the training loss whereby the number of epochs increases as the value of the training loss decreases until it reached the minimum value of 0.2069 when the accuracy of 94.48% was reported.

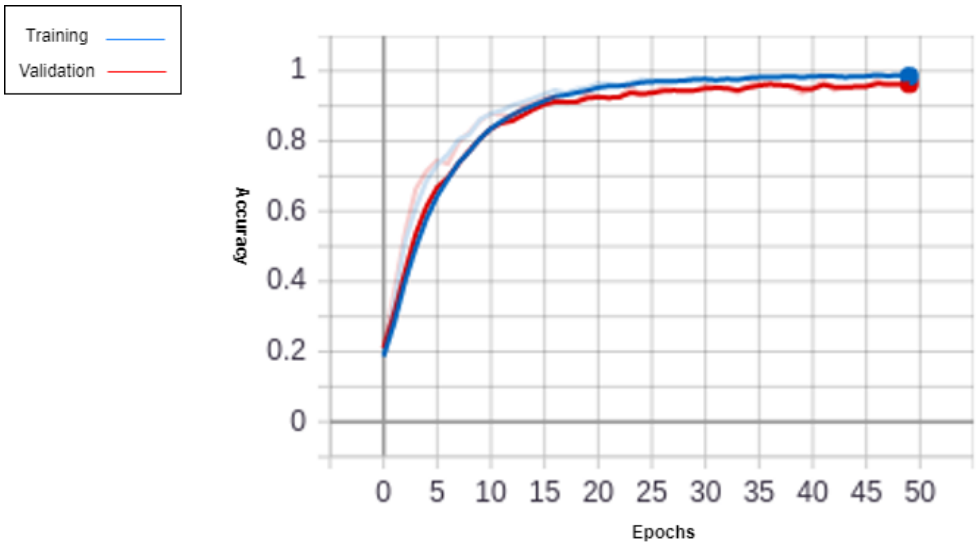


Figure 4.3: Training and Validation Accuracy: Emotion Recognition Model

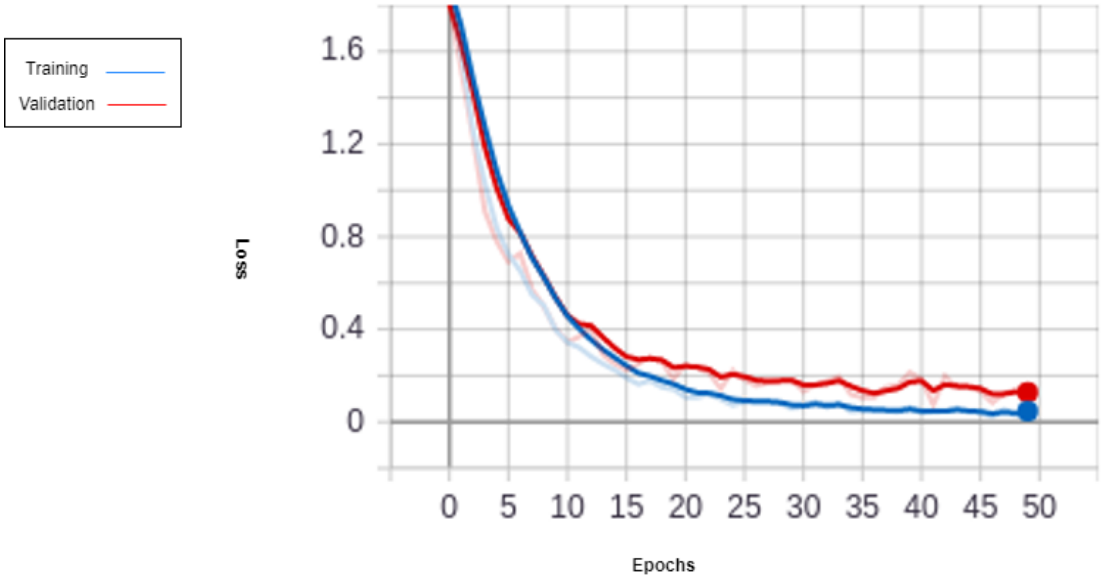


Figure 4.4: Training and Validation Loss: Emotion Recognition Model

As presented in the confusion matrix in Figure 4.5, 17 samples out of 424 samples annotated as anger, were misclassified as other emotions, representing a 4% classification error. Anger was highly misclassified as digust emotion, accounting for 2.6% of the misclassified samples. This shows that anger and disgust might have been mislabeled at some points. The model correctly classified 839 samples annotated as disgust. However, 5 samples of disgust emotions were classified incorrectly, with sadness being classified as disgust mostly, giving the model a 0.6% misclassification rate. This shows a very minimal misclassification rate for disgust.

The model misclassified 78 out of 933 samples annotated as fear emotion, leading to an approximate 8.4% misclassification rate for fear emotion. The misclassification rate for fear emotions by the classifier is a bit high, compared to other misclassifi-

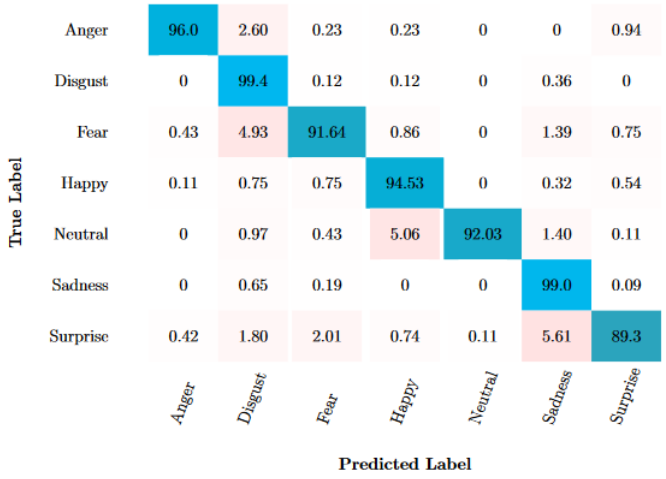


Figure 4.5: Confusion matrix depicting validation results reported by Emotion Recognition Model (values in %'s)

cations. This might be due to mislabelling of the datasets at some points, with fear being highly misclassified as disgust emotion (4.93%).

Out of 933 samples annotated as happy emotions, the model misclassified 51 as other emotions, leading to an approximate 5.5% misclassification rate. The classification error of happy emotion samples is a bit minimal compared to anger and fear emotions. The model further misclassified 74 neutral emotion samples from 929, resulting in about 8% misclassification rate. Error reported on neutral samples was also a bit high and might be due to similar features or mislabelling of samples, as it was highly misclassified as happy emotion (5% of the total misclassifications). With the 1076 samples annotated as sadness emotion, there was a misclassification of 10 samples, resulting in a 0.9% misclassification rate. This is an indication of a very low misclassification rate for sadness samples, as a minimal number of samples from the other emotions were misclassified as sadness emotions, this is an appreciable result. 944 surprise emotion samples reported a misclassification of 101, resulting in about 10.7% misclassification rate. The misclassification for surprise emotion is high compared with fear, neutral emotions that were highly misclassified. We observed that samples of surprise were highly misclassified as sadness (5.6%) and fear accounting for 2% of the total misclassifications. This indicates that these samples have similar features.

The overall results reported an approximately 5% average misclassification over the entire test dataset. Given the minimal misclassification rate, this study suggests that the proposed model has shown an appreciable accuracy when compared with other similar studies [108, 255, 216, 256, 257]. Table 4.2 shows the validation results of the baseline experiment.

Table 4.2: Validation Results: Emotion Recognition Model

Emotions	Number of Samples	Precision	Recall	F-1 Score
Anger	424	0.98	0.96	0.97
Disgust	844	0.90	0.99	0.94
Fear	933	0.96	0.92	0.94
Happy	933	0.93	0.95	0.94
Neutral	929	1.00	0.92	0.96
Sadness	1076	0.90	0.99	0.95
Surprise	944	0.98	0.93	0.93
Total	6083			
Macro Average		0.95	0.95	0.95

4.5.2 Affective State Classifier

For the affective state classification experiment, the following stages were discussed: (i) data pre-processing and feature extraction using pre-trained VGG-19, (ii) Classification of extracted features using BiLSTM classifier; and (iii) Optimization using the validation set. The framework for affective states classification is as shown in Figure 4.6.

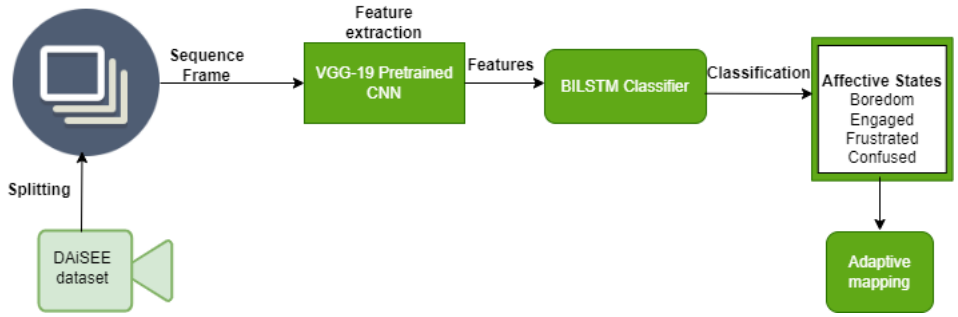


Figure 4.6: Affective State Classification Framework

(i) Data Pre-processing and Feature Extraction: DAiSEE

The pre-trained VGG-19 CNN accepts the input video sequences from DAiSEE dataset. The sample sequences were split into different frames of images. Of the 39,185 sample sequences used for the experiment, 34,880 samples were for training and validation of the model. The original $480 \times 640 \times 3$ frames were resized to $240 \times 320 \times 3$ frames, which is the input size required for the VGG-19 network.

The images forward propagate through the network, and the propagation stops at the last max-pool layer, in which the value of the output shape was 512, which represents the discriminative feature vector. The network, therefore, produced a 2D vector of size 34,880 x 512. The extracted features were passed into the Bi-LSTM network, where the classification is done.

(ii) Training and Validation of Affective State Classification Model

The extracted discriminative features were reshaped to a 3D array comprising the shape, sample size, sequence length, and input dimension. The input to the BiLSTM classifier is a sequence of frames comprising a 34,880 × 512 number of features from each frame. These sequence frames were divided into 20,946 and 13,934 samples for training and validation, which were further passed to the classifier using a batch size of 143 and a maximum epoch set to 50. Binary cross-entropy was used to identify the loss function, and the sigmoid function was used to compute the probability of the output layer.

(iii) Validation result: Affective State Classification Model

This subsection presents the validation results reported during training and validation of the affective state classifier using DAiSEE affective state samples. The BiLSTM network classifies the input feature, considering both the temporal and spatial aspects of the sequence samples, and validates the classification using validation datasets of 13,934 samples. An accuracy of 89% and an F-1 score of 0.89 was reported on the validation dataset. Table 4.3 shows the average precision, recall and the F-1 score that the model reported.

Table 4.3: The Results of the Validation Experiment on DAiSEE dataset

Emotions	**Number of Samples**	**Precision**	**Recall**	**F-1 Score**
Boredom	3338	0.87	0.83	0.85
Confusion	3402	0.77	0.99	0.87
Engagement	3332	0.96	0.74	0.84
Frustrated	3862	0.99	0.99	0.99
Macro Average				0.89
Total Sample	13934			

The training and validation accuracy graph, and the training and validation loss graph, are shown in Figures 4.7 and 4.8, respectively. The graph in Figure 4.7 shows how the training progress related to the number of epochs. During training, the accuracy reached its maximum at the 48th epoch, after which there was no improvement. Further training would result in an over-fitting of the model, therefore the training was stopped. Engagement, confusion, frustration and boredom were the affective states classified. The graph in Figure 4.8 shows the training loss. As the number of epochs increases, the value of the training loss decreases until it reaches the minimum value of 0.2050. The training reported an accuracy of 89% on the DAiSEE validation dataset, which we considered an appreciable level of accuracy.

Out of 3338 samples of DAiSEE validation set annotated as boredom, 584 boredom feature samples were incorrectly classified as other affective states, resulting in a 17.5% misclassification rate. We observe that boredom samples were highly misclassified as confused affective states, accounting for 14.1% of the total misclassification.

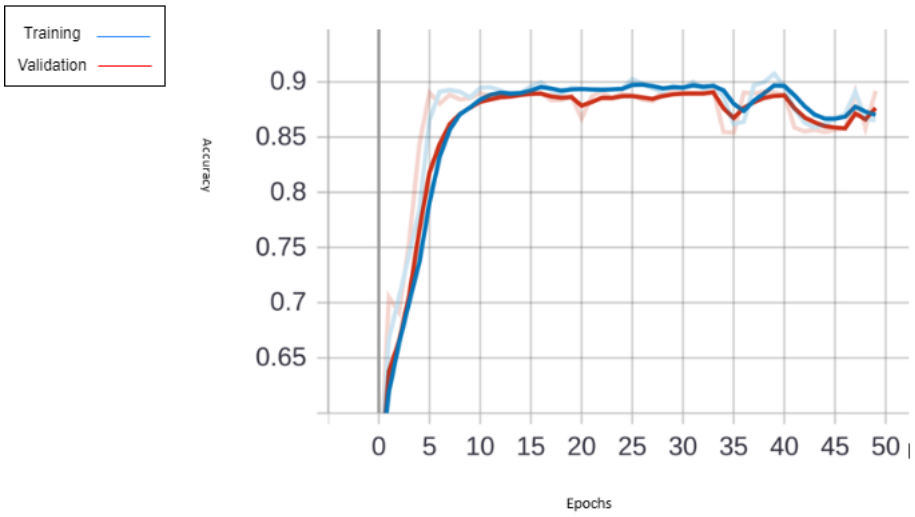


Figure 4.7: Training and Validation Accuracy: Affective State Model

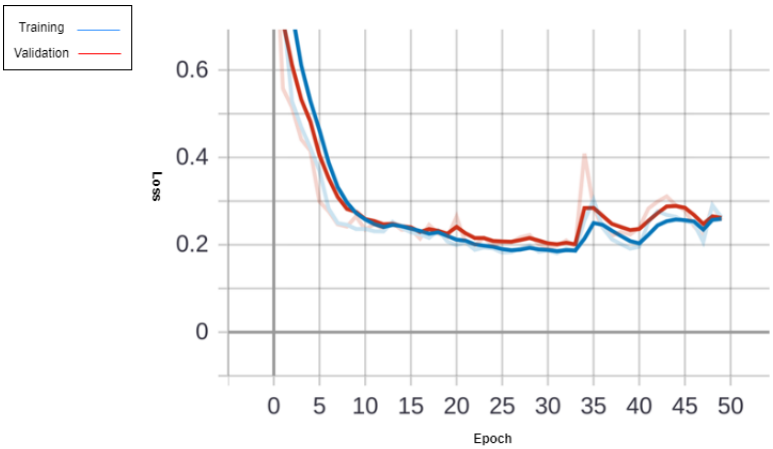


Figure 4.8: Training and Validation Loss : Affective State Model

This might be due to similar features in boredom and confusion samples. The model reported a 0.6% misclassification rate while classifying 3402 samples annotated as confused affective states. The misclassification was from frustrated affective state samples being misclassified as a confused affective state, and these samples account for the whole 0.6% misclassification reported by the model. This revealed that other affective states were not at any point misclassified as confused, which is an appreciable report. Of the 3332 samples annotated as engagement affective state, the model incorrectly classified 884 feature samples, resulting in a 25.6% misclassification rate. This is a very high misclassification report by the model. This study observed that boredom and confused affective state samples were the samples that were highly misclassified as engagement affective states and accounted for the 25.7% reported. This resulted in a recall report, which is the correctly predicted samples out of the total samples of engagement affective state as 74% by the model.

The classifier also incorrectly classified 23 of the 3862 samples annotated as frustrated affective states as other affective states, resulting in a 0.6% error. From

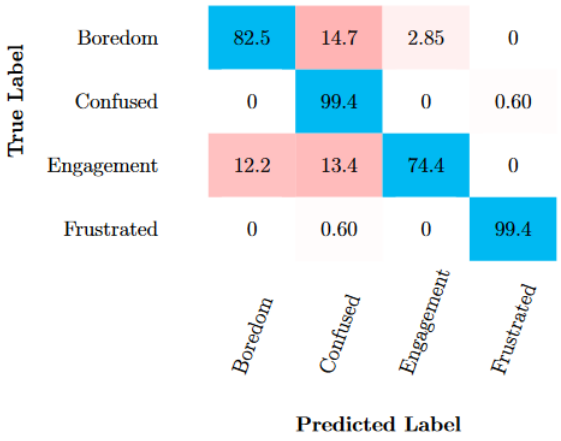


Figure 4.9: Confusion Matrix: Validation Result on DAiSEE (values in %'s

the report, this study observed that confused affective state was the sample that was misclassified as frustrated affective state, which is the only affective state that has accounted for the misclassification reported. The error rate for frustrated affective state was minimal from the report presented by the model and gave an appreciable report of 99% recall and an F-1 score of 99%. The confusion matrix depicting the correct and misclassifications on the validation set is presented in Figure 4.9. Overall, the model reported an approximately 11% misclassification rate.

This study revealed that the classifier has shown a significant level of accuracy, given the low misclassification rate and comparable with other work [19, 219, 218, 220].

4.6 Emotions/Affective States to Learning Affect: Initial Mapping

Based on the literature review, the mapping of estimated emotion and affective states to learning affect is presented in this section. Mapping estimated emotions and affective states to relevant learning impacts aids in understanding the experiences of students throughout a teaching and learning session. Based on the existing literature in [25, 26, 12, 27], this study proposed an initial mapping between emotions and learning affect and this is discussed in 4.6.1. The authors also mapped affective states onto learning affect, and this is discussed in 4.6.2. The mapping between affective states and learning affect was motivated by work presented in [28, 29, 30, 31, 32, 33, 34, 35, 36].

4.6.1 Emotions to Learning Affects

Following the classification, various sources were used to map the anger, disgust, fear, happiness, sadness, surprise and neutral emotions into the positive, negative and neutral learning affect. The mapping information provided by Sathik et al.[25], Kapoor et al.[26], Pan et al.[12], and Zakka et al.[27] highlight some sources. Similar to studies by [25, 26, 27, 12], this study considers anger, disgust and sadness as

negative learning affect. Contrary to that, fear, happy and surprise are considered to be positive learning affects [25, 26, 27]. Even though this study agrees with happy and surprise as positive learning affect interpretations, this study considers fear undefined until further verification. This study undefined neutral emotion as it is considered not to be a positive or a negative learning affect. This deduction was based on [258], where the authors explain that a person experiences neutral emotion when they do not respond to situations but remain in their original state. Gasper et al.[259] stated that neutral emotion is independent of positive and negative affect. Gasper et al. [259] also defined neutral affect as a feeling of indifference to anything in particular and a lack of strong preference. This literature served as a starting point for the mapping performed in this study, and the experiments proposed in this study were designed to consolidate the initial mappings. Table 4.4 summarizes the initial mappings based on the literature.

Table 4.4: Mapping Emotions to Learning Affect based on [25, 26, 12, 27]

,

	Anger	Disgust	Fear	Happy	Sadness	Surprise	Neutral
Sathik et al.[25]	Negative	-	Positive	-	Negative	Positive	-
Kapoor et al.[26]	Negative	Negative	Positive	Positive	Negative	Positive	-
Pan et al.[12]	Positive	Negative	-	-	-	Positive	Positive
Zakka et al.[27]	Negative	Negative	Positive	Positive	Negative	Positive	Positive

4.6.2 Affective States to Learning Affects

Students' affective states might be mapped onto positive, negative, and neutral learning effects [170]. Several studies have been carried out in this regard. To obtain initial information about the categories into which the affective state is mapped, existing literature was considered. These include Daschmann et al.[28], D'Mello et al.[29, 30], Stein et al.[31], Nakamura et al.[32], Csikszentmihalyi et al.[34, 33], Shernoff et al.[35] and Guo et al.[36]. Boredom, confusion and frustration were mapped as a negative affective state by some of these authors, while engagement was considered a positive learning affect. Students who are bored are more likely to have negative outcomes such as poor grades, absenteeism, and finally dropping out of school, indicating a negative affect on learning [28]. Similarly, students experience boredom when the lecture does not interest them or when they cannot complete any of the tasks that are part of the learning process, resulting in a negative learning impact [29]. Confusion is frequently induced by a lack of understanding of an issue, particularly when students face a problem during their learning activities or are unsure about the knowledge being disseminated. Therefore, confusion was considered a sign of negative learning affect [30]. When students need clarification, especially

when they need to understand or comprehend content, they can become frustrated and make repeated mistakes. As a result, frustration was associated with negative learning affect [29, 31]. An intense feeling of flow-like engagement can be linked with a positive learning affect, especially when the learning outcomes are clear and the lecture content is well comprehended [34, 260]. Table 4.5 summarizes the mapping of affective states to learning affect based on literature.

Table 4.5: Mapping Affective states to Learning Affect based on [28, 29, 30, 31, 32, 33, 34, 35, 36]

	Boredom	Confused	Frustration	Engagement
Daschmann et al. [28]	Negative	-	-	-
D'Mello et al. [29, 30]	Negative	Negative	Negative	Positive
Stein et al. [31]	-	-	Negative	-
Nakamura et al. [32]	-	-	-	Positive
Csikszentmihalyi et al. [33, 34]	-	-	-	Positive
Shernoff et al. [35]	-	-	-	Positive
Guo et al. [36]	-	-	-	Positive

4.6.3 Validating Mappings

Due to disagreement between different authors with respect to the mappings between emotions and learning affect, the current study proposed a two-stage experimental setup where the existing mappings between affective states and learning affect can be utilised to validate the mappings between emotions and learning affects (especially the undefined ones). The first stage of this experimental setup includes passing each sample from DAiSEE dataset (annotated with affective states) through the emotion classifier to estimate the emotion label. Mapped learning affect based on the estimated emotion label (utilizing emotion to LA mapping) will be cross-validated with that of mapped learning affect based on the annotated affective state of the sample (utilizing affective state to LA mapping). The framework for the experimental setup is illustrated in Figure 4.10.

The emotion classifier model that was trained with DISFA+ dataset was tested on 2,727 sequence samples from DAiSEE dataset labelled with affective states of engaged, boredom, frustration and confusion. The DAiSEE sample sequence dataset of size $480 \times 640 \times 3$ was reshaped to $240 \times 320 \times 3$ and passed through ResNet-50 pre-trained network for feature extraction, producing a feature size of 2727×2048. The BiLSTM emotion estimator would then map the affective state samples of DAiSEE with the basic emotion from DISFA+. The predicted emotion labels for the affective state samples and the estimated learning affect were compared to the learning affect estimate using basic emotion samples. The mapping results are provided in Table 4.6.

The trained emotion classifier, when tested on samples annotated with engaged affective states, reported the presence of happy and surprise emotions. The affective state of engagement according to Nakamura et al.[32], Csikszentmihalyi et al.[33, 34],

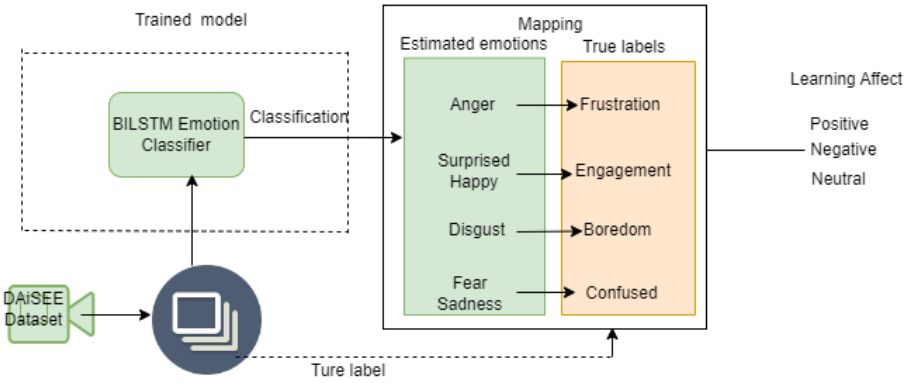


Figure 4.10: Framework for Validating Mappings

Table 4.6: Mapping Emotions to Affective State to Cross-Validate Estimated Learning Affect

DAiSEE Labels	**Sample size**	**Estimated learning affect by literature**	**Mapping by emotion estimator**	**Estimated learning affect by current study**
Engaged	691	Positive	Happy, Surprise	Positive
Boredom	683	Negative	Disgust	Negative
Confused	923	Negative	Fear, Sadness	Negative
Frustrated	430	Negative	Anger	Negative
Total Samples	2727			

Shernoff et al.[35] and Guo et al.[36] were considered as positive states, which in turn validates both happy and surprise emotions to be positive learning states. This was in line with work by [26, 27] who mapped happy to be positive and work by [214, 26, 12, 27] who mapped surprise to a positive state. This reaffirmed the initial mapping of happy and surprise emotions as a positive learning effect. This study therefore agrees and concludes that happy, surprise, and engaged affective states shown during learning by students are indications of a positive learning affect.

Samples annotated with the confused affective state were reported to be depicting sadness and fear emotions by the emotion classifier. The affective state of confused was considered a negative affective state by Graesser et al.[80] which in turn maps sadness and fear as a negative learning affect. Mapping of sadness to negative learning affect was in line with work by [25, 26, 27]. However, fear was considered a positive learning affect by [25, 26, 27]. This prompted further investigations to confirm the mapping of fear. This study, therefore, agrees that sadness and confused affective states are negative learning affects, but still investigated fear emotion further. Section 5.6 discussed the survey that was carried out along with the specific questionnaires administered to the participants in this aspect.

Samples annotated with the boredom affective state, which was considered to be a negative learning affect by [28, 29], were reported to be depicting disgust emotion. This maps disgust emotion being a negative learning affect and this was in line with the work by [26, 12, 27]. This study also agrees with this mapping of boredom and disgust as depicting a negative learning affect. This implied that, when students are showing a sign of boredom or a disgust emotion, there is no proper understanding of the subject matter, and this should send a signal to the teacher, that the student is not grasping what he or she is teaching.

Anger emotion was picked up when samples annotated as being frustrated affective states were passed through the emotion estimator. Affective states of frustrated were considered as negative learning affect [29, 31], which in turn maps anger emotion to be a negative learning affect. This was in line with the work by [25, 26, 12, 27], where the authors reported anger emotion to be a negative learning affect. This study agrees with this and concludes that anger and frustrated affective states are indications of poor learning, and eventually, a negative learning affect. Negative learning affect in turn might lead to eventual failure at the end of the session. The second stage includes a survey with participants to re-evaluate the fear emotion and its mapping to a learning affect.

4.7 Feedback Generation

Feedback is crucial for effective teaching and learning either in a traditional classroom setting or in an e-learning setting. Feedback reveals how well students are supporting themselves in achieving learning objectives and improving self-regulation abilities[261]. Without feedback, the learning process would not be complete since it would not be possible for teachers or students to assess how well they are doing regarding the learning goals and to determine if they are being met or exceeded.This study proposed two feedback mechanisms; narrative reports (on overall estimates for a student or a lecturer) and Q&A (for individualised feedback). Narrative reports using passage templates and estimates from emotion and affective states classifier estimated learning affect, lecture video annotations (i.e., topic, length of the lecture, etc) will be formulated. Whereas for Q&A-based feedback, this study adopted a fine-tuned BERT-Q&A bot. The BERT-Q&A bot accepts reasoning questions and passages (which are narrative reports in this study). The formulated questions will be in line with and related to the teaching and learning aspects, which warrant an answer with reasoning embedded by the BERT-Q&A bot. The details of the report generation and the BERT-Q&A bot modules are described in subsections 4.7.1 and 4.7.2 respectively.

4.7.1 Narrative Report Generation

In the narrative report module, the task is to automatically generate a narrative report from passage templates and estimates from emotion and affective states classifier, estimated learning affect, and lecture video annotations (i.e., topic, length of the lecture, etc). The narrative report will act as the input to the next phase of

feedback based on the BERT-Q&A bot. The narrative reports fall under two categories: single sample input (for students), and a group sample input (for teachers). Group sample input will cater for teachers who want to get feedback based on the learning experienced by all students during an e-learning session.

For the report, we manually curated a set of sentences as a template. The template was built by examining the output generated from the classifiers and the learning affect outcome. Two sets of templates were created. One is to accept input from the single/individual sample model and the other one is to accept input from the group samples model. Sentences that can reflect on the length of the video lecture, lecture labels, the estimated emotion and affective state and learning affect were curated for both single and group sample models. We adopt the PASS method of data-to-text in [201] for template-based report generation. The template is created to have empty slots that needed to be filled to complete the report. This is carried out by the template filler module as described in [201]. Each empty slots has a tag in the template database, e.g $< happy >, < bored >$, etc. The template filler module uses these tags to find the corresponding information in the model output data, and then fills the empty slots with the relevant data values. The filled-in template is returned as the output of the report generation module.

Single report template sample: During the $< video \quad length >$min video lecture on $< video \quad label >$ that you watched, you have mostly expressed $< highest \quad emotion \quad type >$ emotion and $< highest \quad affective \quad state \quad type >$ affective state. In addition to this, $< second \quad highest \quad emotion \quad type >$ emotion and $< second \quad highest \quad affective \quad state \quad type >$ affective state were exhibited. Based on the emotions and affective states that were mostly expressed during $< video \quad annotation >$, the system predicts a $< highest \quad Learning \quad Affect >$ learning affect. You are advised to consider the above feedback and adapt your learning session accordingly for future video lectures.

Group report template sample During the $< video \quad length >$min video lecture on $< video \quad label >$, $< highest \quad class \quad aggregate >$% of learners expressed $< highest \quad emotion \quad type >$ emotion and $< highest \quad affective \quad state >$ affective state. In addition to this, another $< second \quad highest \quad class \quad aggregate >$% of learners expressed $< second \quad highest \quad emotion \quad type >$ emotion and $< second \quad highest \quad affective \quad state >$ affective state. Based on the emotion expressed by most learners, the system predicts a $< highest \quad Learning \quad Affect >$ learning affect. The teacher is advised to consider the above feedback for future video lectures.

The report generator outputs the completed template, which we label as narrative feedback reports. This output is then forwarded to the BERT-Q&A bot as the input passage to generate answers for the questions posed by either teacher or student for further feedback.

The samples of the input questions that were formulated in this study are as follows (This is just a sample of questions and it depends on the specific needs of students/learner and teachers, more questions can be embedded into the list.):

Pre-defined Input Questions

Set of predefined questions to be used along with the narrative report generated from the single report include:

1. Which emotion do I exhibit the most?
2. Which affective state do I exhibit the most?
3. During the first 2 minutes of the lecture, which emotion and affective state did I exhibit?
4. During the first half of the lecture, which learning affect was estimated for me?
5. Which emotion did I exhibit towards the end of the lecture?
6. Did I understand the lecture?
7. Which topic do I seem to not comprehend well?
8. How can I improve?
9. What can the teacher do to help me improve?
10. Which learning affect do I experience?

Set of the predefined questions to be used along with the narrative report generated from the group report include:

1. What was the most frequent emotion exhibited by the class?
2. What was the most frequent affective state exhibited by the class?
3. During the first 5 minutes of the lecture, what were the highest emotion and affective states exhibited by the students?
4. What was the overall learning affect?
5. What can the teacher do to assist the students?
6. Were the students distracted during the video lecture?
7. During the first half of the lecture, which learning affect was estimated for the class?
8. Were the students frustrated during the lecture?
9. Did the students understand the lecture?
10. Which learning affect was estimated for the entire class?
11. How could the teacher rate the content of the lecture?

4.7.2 Question & Answer Module : BERT-Q&A bot

In the Q&A module, the user, who is either the teacher or the student, asks questions through the BERT-Q&A bot. These questions were asked along with the passage generated from the emotion and affective state estimator, learning affects mapping, the video label, and its length to provide relevant answers to the questions posed. The BERT-Q&A bot model receives the narrative report and a question from the

user (teacher/student). The questions would be chosen from the pre-defined set as described previously, and these questions are expected to help get reasoning-based answers by the BERT-Q&A bot. The set of questions has the potential to be expanded in the future. As the input level, questions and narrative report are fed in as a single package sequence, with the question as the A embedding, and the passage as the B embedding. A start vector $S \in R^H$ and an end vector $E \in R^H$ are introduced during fine-tuning [24]. Dot products between Ti and S are used to calculate the probability that, word i will begin the answer span, which is then multiplied by a softmax across each word in the paragraph and outputted as answers:

$$P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}} \tag{4.1}$$

The Q&A module returns an answer along with the start/end token index for the answer and the start/end token scores. The scores can be used to rank answers if we are searching for answers to the same questions in multiple passages. The value of the maximum length can not exceed 512. If the length of the question and a passage with special tokens is larger than the maximum length, the function will truncate the extra portion. The answers provided through the input are expected to serve as feedback for teachers and students.

4.8 Live Testing

To evaluate and explore the feedback aspects of the proposed framework, a live test was carried out with real-life samples from students in a laboratory setting. Fifty-one students from the Department of Computer Science and Applied Mathematics (CSAM) at the University of the Witwatersrand volunteered to test the proposed system. Each participants was made to watch a 10-minute lecture video on a topic of interest in the area of computer science, using a computer system with a webcam. The webcam was activated as soon as the learning session started in order to collect the student's facial images. Input sequences were pre-processed (resized and reshaped) and forwarded to the trained CNN-BiLSTM model for emotion and affective state classification. The highest estimated emotions and affective states were displayed by the proposed model at the end of the lecture session. The learning affects are estimated and passed to the template-based report generation module, thereby populating the template. The template accepts estimated emotion and affective state, estimated learning affect, and the video annotations (label/length etc.). This process is carried out on a group basis at the teacher's end, whereby a group-level emotion, affective states and learning affect populate the template. The populated templates and questions from the user (teacher/student)are then forwarded to the BERT-Q&A bot.

4.8.1 Ethical Considerations

This study required human participants to evaluate the proposed framework. Therefore, an application for ethical clearance was sent to the Wits University Human

Research Ethics Committee (non-medical) for approval. Each participant would be required to view a lecture video on a topic of interest in the area of computer science and their emotions/affective states will be estimated and used to provide feedback on the student experience.

Ethics application included a full explanation of what, why, and how the work will be done, as well as a completed application form. A sample approval for this study by the head of the school and the ethics clearance certificate are shown in appendixes D and E, respectively. The researcher also prepared a set of documents that participants are expected to fill out to complete a survey on the suggested model. These documents are explained as follows.

1. **Application form**: The categories of work risks are listed in the application form that the researcher filled out. This study was classified as having minimal risks. The participants were not chosen as experts according to the form. The researcher wrote to the committee that the study participants did not belong to any vulnerable groups. The study goals and objectives were also indicated in the application form, along with the mode of the survey to be conducted and the categories of participants, which include all registered students of the School of Computer Science and Applied Mathematics. The application also includes other documents that were provided to the participants.
2. **Participant's information sheet:** This document introduced the researchers and explained the study to the participants in detail, including how they would participate in the live testing of the proposed model. Participants were informed that they were not under any obligation to participate and that participation was entirely voluntary. The participants' form is provided in Appendix A.
3. **Consent forms**: This form provides information to a participant, as well as his or her consent to participate or not participate in the survey and a promise of anonymity. The participant's signature is also included in the form, indicating his or her willingness to participate in the survey. A consent form is provided in Appendix B.
4. **Questionnaire**: The researcher prepared a questionnaire to conduct the survey to evaluate the proposed model in relation to the basic emotions and affective states of the students. A qualitative analysis of the study was carried out since the data used for the survey are qualitative. Wolter et al.[262] described qualitative data as data from direct fieldwork observations, open-ended in-depth interviews, and written documents. This is analyzed in a qualitative research framework. Naturalistic inquiry is used by qualitative researchers who explore real-world contexts inductively to provide rich narrative descriptions and case studies. Patterns and themes emerge from inductive examination across cases and the result of qualitative research.

The formulated questions were geared to extract information about the basic emotions experienced and are divided into the following categories: Questions 1 to 3 were formulated to evaluate the proposed model. This is intended to compare the responses of the participants with the output of the model in terms of reasoning.

Question 1 tries to capture the emotion experienced by the participants during the lecture session, apart from what the model has estimated. Question 2 tries to capture the learning affect experienced by the participant. This information is used to compare the initial mappings proposed in this study. Question 3 is about the rating of the lecture with respect to the mode of presentation. This is to compare the learning affect and the presentation method of the lecture to see if the presentation method has an effect on the learning effect of the students. Questions 4 to 9 were formulated to capture the general view of the proposed research and its importance. The potential responses included "yes", "no", or "I don't know". This is to estimate the impact of the proposed study. Specifically, question 4 was formulated to find out from the participants if facial emotion is an acceptable method to evaluate a student during an online learning process. Question 5 was intended to confirm whether the estimated learning affect of students using their facial emotions can improve the e-learning domain. Question 6 tries capturing the feedback around the proposed method around the level of comprehension experienced by a student. Question 7 tries to capture participants' feedback with respect to their opinion on the proposed study and its capability in helping teachers adequately in assessing students during an online teaching session. Question 8 was intended to obtain suggestions for other criteria that can be used to estimate the learning affect of students in an online setting. Question 9 tries to establish whether there is room for improvement on the proposed system followed by question 10, which tries to capture suggestions for improving the study. As a way to determine whether the students' learning experiences align with the model's predictions, Question 11 was asked. The questionnaire is provided in Appendix C.

To perform statistical significance on the data collected from the survey, the study adopted Stata statistical analysis software. This tool was used to conduct Fisher's exact test on the different variables that were observed. Fisher's exact test is employed to determine whether there is a significant relationship between two categorical variables [263, 264].

4.9 Chapter Summary

Reasoning about the learning affects of students during e-learning using the students' facial cues, narrative reports, and question-answers is considered a missing link that this study seeks to solve. The solution to this problem is expected to increase teachers' information about their teaching and students' information about their possible learning performance. Feedback about the learning affect will not only improve the academic performance of students but will also provide a solid learning foundation. Therefore, the application of an appropriate deep-learning approach will in no small way provide the required solution.

The general techniques for achieving the expected solutions were provided in this chapter, with the goal of introducing a system that estimates a student's learning affect, with questions and answers incorporated to provide feedback to teachers and students. The problems that necessitated this study were discussed. These issues identified include the absence of reasoning using questions and answers to estimate students' learning affect as feedback generation for teachers and students.

The proposed framework presented comprised lecture input from the teacher, video facial input from students, feature extractors (ResNet-50 and VGG-19 pre-trained CNN), emotion and affective state classifiers (BiLSTMs), the learning affect module, report generation module, feedback templates, and the BERT modules for feedback.

The data and its format for the experiments, as well as the different experiments carried out towards the achievement of the proposed study, were described in this chapter. These experiments include training the emotion and affective state classifiers using an existing classifier model and initial mappings from estimated emotions and affective states to learning affects. These mappings were further validated through another experiment (mapping emotion to affective states).

The chapter also provided an overview of the feedback mechanism, including both narrative reports and the Q&A module. The live test experiment and ethical considerations, including the various documents that were processed to obtain clearance, were also discussed. Results obtained through these experiments carried out are presented and discussed in chapter 5.

The estimation of students' learning carried out using their facial emotions and affective state analysis with narrative reports and question and answer to reason about the learning affect of students is expected to improve the e-learning domain. Getting feedback on students' learning is crucial for both students and the teacher. Therefore, we believe that providing a system to facilitate feedback is expected to contribute to the e-learning sector since there are no physical contacts, unlike in the classroom setting.

Chapter 5

Results and Discussion

5.1 Introduction

The proposed spatio-temporal reasoning framework, as described in Chapter 4, was evaluated for its correctness and effectiveness. The results obtained are presented and analysed in this chapter. The chapter comprised eight sections. Section 5.2 reviews the data samples used for the test experiments, followed by Section 5.3, which presents the test results obtained by the proposed emotion classifier and affective state classifier. Experimental results from adaptive mapping are presented in Section 5.4, while the overview of the feedback received by the learners/teacher during the online session and after the lecture is discussed in Section 5.5. Live test experimental results are presented in 5.6, followed by a discussion on the experiments in Section 5.7. The summary of this chapter is presented in Section 5.8.

5.2 Data Samples for Testing the Proposed Framework

For the inter-database test experiments, we used 2,274 data samples from DISFA+ that were not part of the training and validation experiments. A feature vector of size $2,274 \times 2,048$ reshaped into a 3D array was used as the input to the emotion classifier. On the other hand, a sample size of 4,305 sequences was used from DAiSEE dataset to test the affective state classifier. A feature size of 4305×512 was reshaped into a 3D array by the feature extractor before passing them to the affective state classifier. For live test experiments, 51 video samples depicting learners' facial data during an e-learning session were captured in laboratory settings.

5.3 Experimental Results

5.3.1 Test Results: Emotion Classifier

The test experiments carried out on the trained emotion classifier reported classification accuracy of 92% on a sample size of 2,274 depicting seven basic emotions i.e, anger, disgust, fear, happy, sadness, surprise, and neutral. An average F-1 score

of 88% and an average precision of 89% were also reported. Table 5.1 provides the emotion classifier test results, and Figure 5.1 depicts the confusion matrix.

Table 5.1: Emotion Classifier: Test Results on DISFA+

Emotions	**Number of Samples**	**Precision**	**Recall**	**F-1 Score**
Anger	434	0.99	0.95	0.97
Disgust	267	0.99	0.95	0.97
Fear	204	0.71	0.62	0.66
Happy	470	0.92	0.95	0.93
Neutral	330	1.00	0.99	1.00
Sadness	253	0.80	0.81	0.81
Surprise	316	0.81	0.92	0.86
Total	2274			
Average		**0.89**	**0.88**	**0.88**

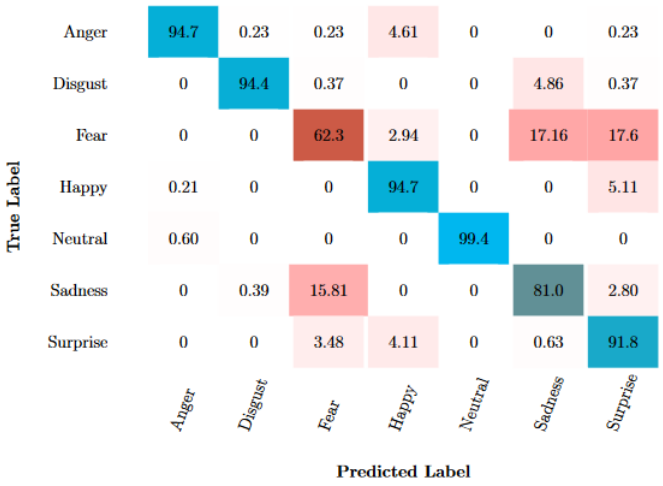


Figure 5.1: Emotion Classifier: Confusion Matrix (values in %'s)

Based on the report provided through the confusion matrix, out of 434 samples of anger emotion, 23 samples were misclassified as other emotions by the classifier, accounting for a 5.3% error. Anger samples were misclassified as disgust, fear, happy and surprise, with the highest misclassification as happy emotion (4.6%). This is a minimal error by the classifier.

With disgust emotion samples, 15 out of 267 were misclassified as other emotions. These emotions, which disgust was misclassified as, are fear, sadness, and surprise, totaling 5.6% misclassification. It was observed that disgust emotion samples were highly misclassified as sadness, which accounts for 4.8% of the total misclassification. The misclassification was minimal in this case too.

For fear emotion, out of 204 samples, 78 were misclassified as sadness, surprise, and happy emotions. The total misclassification was reported to be 38%. Fear was mostly misclassified as sadness and surprise, as depicted in the confusion matrix. This is a very high misclassification rate compared to other reports. 25 samples of happy emotions out of 470 were misclassified. The classifier reported a 5.3% misclassification in this case. Happy emotion samples were misclassified as anger and surprise emotions. The classifier has also reported a minimal error rate in this case.

Two (2) samples of neutral emotion out of 330 were misclassified. Neutral emotion was misclassified as anger emotion, which accounts for the 0.6% misclassification reported. This error is very minimal, and we opined that the classifier has reported an appreciable result on neutral emotion. Of 253 samples of sadness emotions, 48 were misclassified as fear, disgust, sadness, and surprise, reporting an 18.9% misclassification rate. This is a very high error rate reported by the classifier compared to other emotions. Surprise emotion had 23 misclassifications out of 316 samples. These emotions that surprise emotion samples were misclassified as fear, happy and sadness, and this totals 7.2% misclassification. The average misclassification reported during the test experiments by the emotion classifier is 11.5%, we noticed that this error is minimal, and therefore opined that the accuracy of the classifier is appreciable and in line with other literature [108, 255, 216, 256, 257].

5.3.2 Test Results: Affective State Classifier

The CNN-BiLSTM-based affective state classifier reported an accuracy of 86% and an average F-1 score of 87% on a sample size of 4,305. The results were in line with the state-of-the-art results reported by [19, 219, 218, 220]. Table 5.2 provides a summary of the test results, and Figure 5.2 depicts the confusion matrix.

Table 5.2: Affetive State Classifier: Test Results on DAiSEE

Emotions	Number of samples	Precision	Recall	F1-Score
Boredom	782	0.98	0.91	0.94
Confusion	758	0.89	0.81	0.85
Engagement	1362	0.84	0.79	0.81
Frustrated	1403	0.81	0.94	0.87
Total	4305			
Average		**0.88**	**0.86**	**0.87**

From the results presented by the confusion matrix, the classifier reported a misclassification rate of 9.3% for affective state samples of boredom, out of 782 samples that were classified. 73 samples were misclassified as confusion, engagement, and frustration. This error reported was considered minimal by this study. For the classification of confusion samples totalling 758, 146 were misclassified as boredom, engagement, and frustrated. The misclassification accounts for a 19.2% error. We opined that this error was high compared with other affective states and therefore affected the final accuracy of the classifier. Of 1362 samples of engagement affective

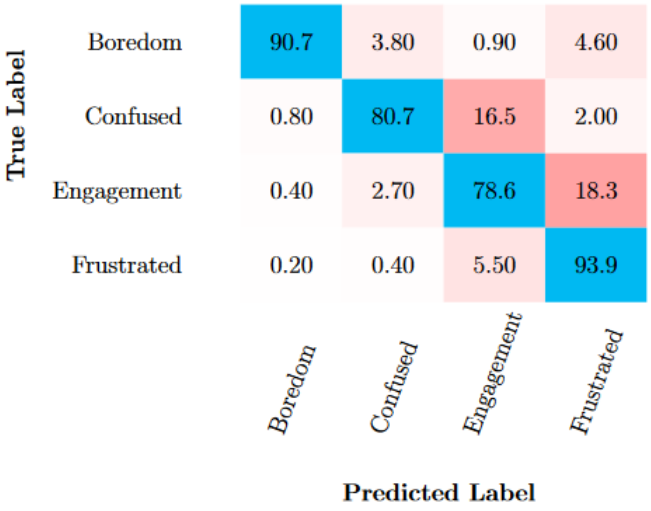


Figure 5.2: Affective State Classifier: Confusion Matrix (values in %'s)

states, 292 were misclassified as boredom, confused and frustrated. This totals 21% misclassification, which is higher than the misclassification reported by confusion affective state. This report will also have some effect on the final accuracy of the model. 1403 samples of frustrated affective states were classified, and 85 samples were misclassified as boredom, confusion, and engagement. This accounts for a 6% misclassification. This model has reported an appreciable error rate for frustrated affective states. The average misclassification rate reported by the affective state classifier during the test experiment was approximately 14%. This error was high and was responsible for the accuracy of the model reported as 86%, but generally, the result is comparable to other similar works.

5.3.3 Comparison of experimental results with related recent studies

Generally, both the emotion and affective state classifiers have produced results that were inline with and comparable to those in the literature. Table 5.3 and 5.4 provide a list of works and how our model compared to them. The comparison metrics include the datasets used, the architecture used, the type of modality used, the mapping criteria, the accuracy of the classification reported, and the availability of the Q&A module.

According to this comparison, no previous study has formulated a complete end-to-end framework for capturing facial cues to generate feedback that incorporates narrative reports and Q&A modules.

Table 5.3: Proposed Framework vs Existing Work using emotion data

Authors	Dataset used	Architecture	Classified Actions	Accuracy Reported	Mapping Criteria	Q&A Module
Hindujah et al.[216]	DISFA+	CNN	Intensities of 12 AUs including AU 1-6,9, 12,15,17,20, 25,26	79%	None	None
Akay et al. [108]	DISFA	Stacked CNN	Detection of twelve AUs namely: AU1, 2, 4, 6, 9, 12, 25, 26	55%	None	None
Hernadez et al.[255]	DISFA	DeepFN	Facial Action unit	69.7%	None	None
Pise et al.[215]	DISFA+	SqueezeNet + TRN	Anger, disgust, fear, happy, sadness, surprise and neutral	92.7%	None	None
Current Study	DISFA+	CNN-BiLSTM	Anger, disgust, fear, happy, sadness, surprise and neutral	**92%**	Positive, Negative, Neutral	Yes

Table 5.4: Proposed Framework vs Existing Work using affective state data

Authors	Dataset Used	Architecture Used	Classified Actions	Accuracy Reported	Mapping criteria	Q& A Module
Liao et al.[218]	DAiSEE	SENET +LSTM+ GALN	Engagement prediction:Very high, high, low, very low engagement	58.84%	None	None
Fwa et al.[220]	DAiSEE	STGAT	Frustration,boredom, engagement and confusion	60.32%	None	None
Abedi et al,[219]	DAiSEE	ResNet + TCN	Engagement level: very low, low, very high, high	63.9%	None	None
Current Study	DAiSEE	CNN-BiLSTM	Boredom,engagement, frustration and confusion	**86%**	Positive, Negative, Neutral	Yes

5.4 Final Mappings

Based on the literature survey, initial mappings were verified. Further experiments via survey questions answered by willing participants and initial mapping were adopted. Based on the experimental results, this study adapts these mappings and

concludes that sadness, disgust, anger, frustration, boredom, and confusion were regarded as negative learning affect, while happy, surprise, fear, and engagement were regarded as positive learning affect. When a learner shows an indifferent emotion that is neither positive nor negative, this is regarded as a neutral learning affect. Table 5.5 shows the mapping adapted as described in this work.

Table 5.5: Adaptive Mapping by this research

Affective states/Emotions	Adaptive Mapping
Sadness, Disgust, Anger, Frustration, Confusion, Boredom	Negative
Happy, Surprise, Fear, Engagement	Positive
Neutral	Neutral

5.5 Feedback Generation

An overview of the feedback received by the learner/teacher during the online session and after the lecture is discussed in this section. During the lecture, the learner is reported on the facial emotion/affective state expressed by him/her. This keeps the learner engaged and provides real-time feedback on the cues expressed by them.

5.5.1 Emotion and Affective State Feedback

Sample frames with estimated emotions and affective states during an online lecture session are shown in Figure 5.3 and Figure 5.4, respectively.

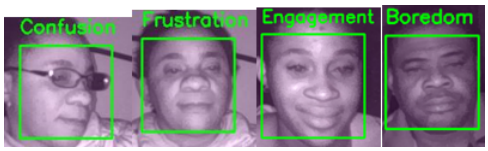


Figure 5.3: Live Testing: Estimated Facial Emotion as Feedback to User

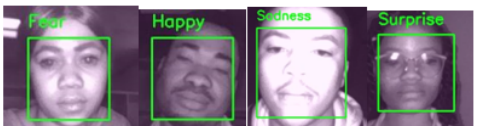


Figure 5.4: Live Testing: Estimated Affective State as Feedback to User

To generate individual and group-level aggregates of emotions and affective states experienced by learners, a statistical module is used, as discussed in the following section.

5.5.2 Individual and Group Aggregates Generation

Highest occurring emotions and affective state estimates were passed to a statistical module (STAT encoding) to evaluate the aggregates. Statistical encoding is an entropy-based compression approach that recognises statistical patterns in the data to be compressed and uses these patterns to perform compression [265]. The statistical module aggregates the estimates of emotions and affective states for a single learner and outputs the learning affect (positive, negative, or neutral) [266] in textual form. For the group level, estimates of emotion and affective state labels are concatenated and aggregated.

5.5.3 Narrative Report Generation Module

To create narrative reports for individual learners and teachers, this study used two sets of templates, which are to be populated using estimates of emotions, affective states, aggregates, estimated learning affect, and video labels. This study adopted the PASS method of data-to-text as in [201] for template-based report generation. In the PASS method, the template is created, having empty slots that must be filled to complete the report. The template filler module carried out this, as described in [201]. Each empty slot has a tag in the template database, e.g., $< happy >, < bored >$, etc. The template filler module uses these tags to find the corresponding information in the model output data and then fills the empty slots with these data. The filled-in template is returned to the report generation module. The empty slots of the template were filled out as a narrative report that serves as the passage for the BERT-Q&A bot. Figure 5.5 shows a sample of the generated report for a single sample, and Figure 5.6 shows a sample of the generated group sample. Both versions will act as input to the BERT-Q&A bot, as explained in 4.7.1.

```
# BERT Question-Answer sample 1 (Single sample narrative report/ passage)

passage="During the ten minute video lecture on machine learning\
 that you watched,you have mostly expressed fear emotion \
 and engaged affective state. In addition to this,\
 happy and confused affective state were exhibited.\
Based on the emotion and affective states that you mostly expressed,\
 the system predicts a positive learning affect.\
You are adviced to consider the above feedback \
and adapt your learning accordingly \
for future video lectures. "
```

Figure 5.5: Narrative Report: Single Sample Report

5.5.4 Question-Answer generation: BERT-Q&A bot

For the Q&A task, the narrative report generated as described in 5.5.3 is plugged into the BERT Q&A module.

```
# BERT Question-Answer sample 1 (Group narrative report/ passage)

passage="During the 10 minutes video lecture on machine learning, \
81.58% of learners expressed fear emotion\
and engaged affective states. In addition to this, another \
60.53% of learners expressed happy emotion and engaged\
 affective state.\
Based on the emotion and affective states expressed by\
 most learners, the system predicts\
positive learning affect. The teacher is adviced to consider\
 the above feedback for future video lectures.  "
```

Figure 5.6: Narrative Report: Group Sample Report

Feedback to learner

For a learner's feedback, the BERT-Q&A accepts the single passage template that has been completed with the single sample details. This passage is input along with questions designed for a single input, and reasoning answers are output by the BERT-Q&A bot. The output of the BERT Q&A bot is shown in Figure 5.7.

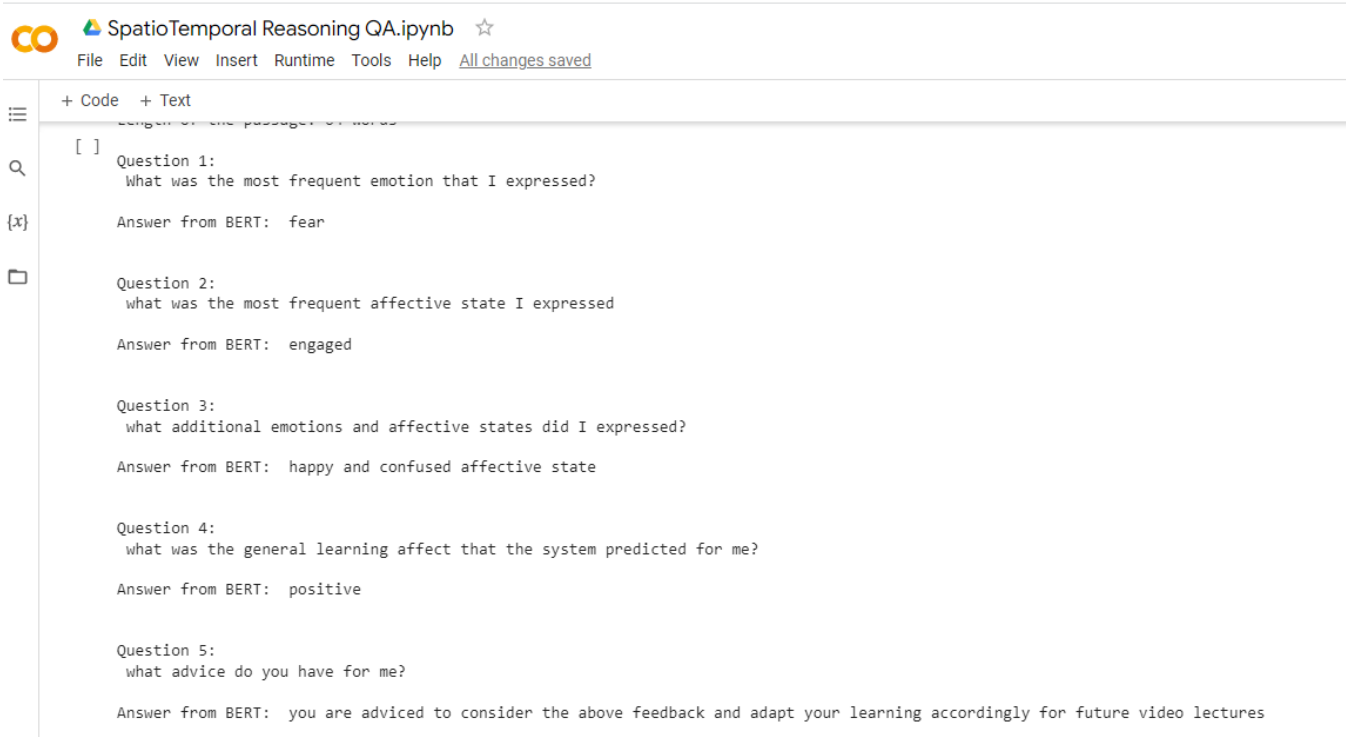


Figure 5.7: BERT-Q & A Bot Output for Single sample QA

Feedback to teacher

For the teacher's feedback, the BERT-Q&A accepts the report from the group sample as the input passage, which was filled with the group sample details. This passage is input along with questions designed for a group, and reasoning answers are produced by the BERT-Q&A bot. The output of the BERT Q&A bot for one group reasoning sample is shown in Figure 5.8.

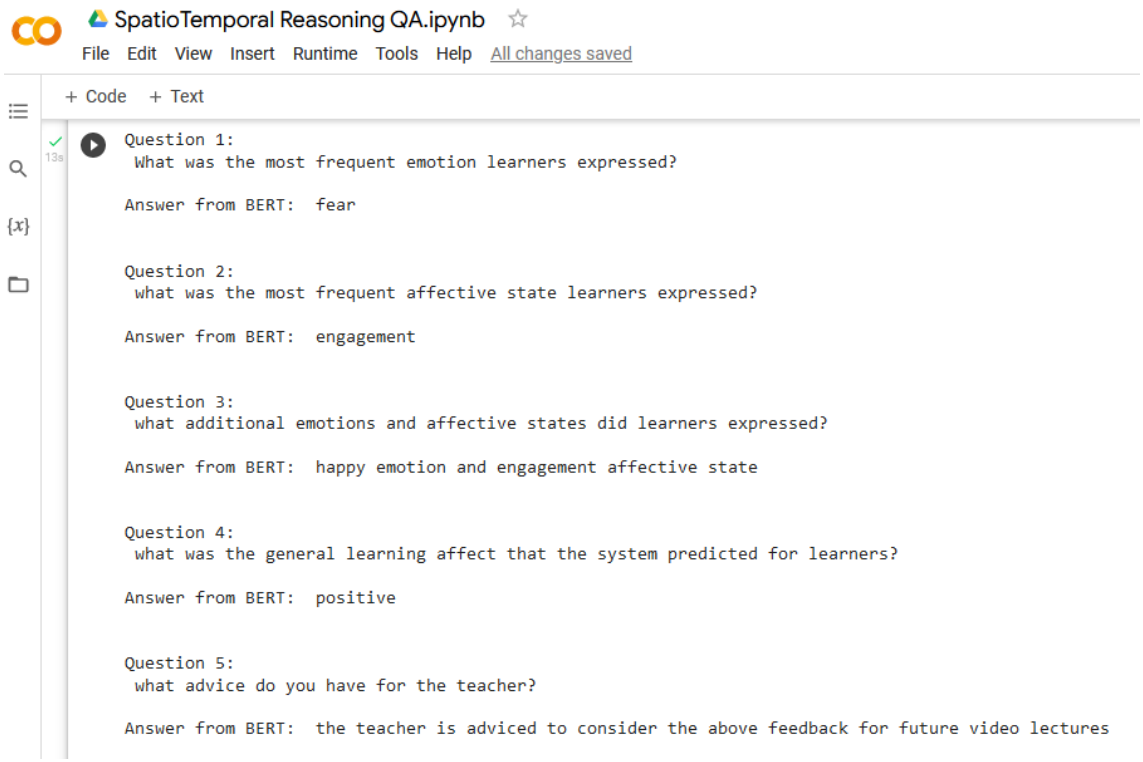


Figure 5.8: BERT-Q & A Bot Output for group Q&A

5.6 Live Test Results

The results of the live test experiment using data samples from learners are reported in this section. The proposed model consisted of a cascade of emotion and affective state classifiers, the learning affect module, the narrative report generation module, and the BERT-QA bot module. The questionnaire provided to the learners was used to validate the estimates at all levels of the framework.

5.6.1 Data Management

All responses provided by respondents on paper questionnaires were captured in Microsoft Excel. The data were exported to Stata version 14 for analysis.

Quantitative Data Analysis

Descriptive statistics were computed for all the quantitative variables. Descriptive statistics are statistics that summarize or describe the characteristics of a data set [267]. Descriptive statistics are made up of three types of measures: measures of central tendency, measures of variability (or spread), and measures of frequency distribution. Measures of central tendency describe the data set's centre (mean, median, and mode). Variability measures describe the data set's dispersion (variance, standard deviation). The frequency distribution measures describe the occurrence of data within the data set (count) [267]. Most variables were categorical and described using frequency distributions and percentages. Tables were used to present visual summaries of the findings. Fisher's exact test was used for bivariate analysis to determine associations between key outcome variables. A significance level of 0.05 (95% level of confidence) was used. P-values of less than 0.05 reflect statistically significant differences.

The study initially used 22 participants for the experiment. By cross-tabulating, the model's emotion estimate with the learners' feelings, a p-value of 0.02 was reported, which is statistically significant and shows that there is an association between the model and the learners' feelings. Therefore, more data samples were collected to confirm the observed significance from the previous samples. The updated sample reports are presented as follows.

5.6.2 Findings

The findings of the proposed model through the live test based on the demographic profiles of the participants and the learner's emotions and affective states experienced during the live test are discussed in this subsection.

Demographic profile of the participants: The demographic profiles of the 51 participants who volunteered to participate in this study ranged between 25 and 50 years of age and consisted of 45% (23) females and 55% (28) males.

learner's emotions and affective states: The different emotions and affective states expressed by participants at different times during the live test of the proposed model were attributed to the dynamic nature of emotions and affective states. This means that emotions and affective states vary over time and space. According to the observations, some emotions and affective states were shown more frequently during learning than others. During the experiment, this study observed that most of the learners frequently expressed fear emotion and an affective state of engagement, although the study agrees that there might be a possibility of misclassification errors at some points. Although emotions and affective states such as happy, neutral, sadness, surprise, frustration, and confusion were expressed less frequently, others, such as anger, disgust, and boredom, were not estimated at any time. With a misclassification rate at 11.5%, it might be beneficial to test the model with a larger number of participants which might bring in all the diverse emotions and affective states.

The proposed model estimated that fear emotion was prevalent in 62.75% (equivalent to 32) of the participants, and the baseline experiment revealed that the proposed model misclassified the samples of fear emotions. During the second exper-

iment conducted to verify the initial mapping, the emotion estimator labelled fear a confused affective state of learning. Authors suspected this to be because the baseline model had a high rate of fear misclassification, as shown in Table 4.2. The previous literature that was surveyed in this study opined that fear is a positive emotion. Participants verified this through the questionnaire as a positive learning affect. Therefore, we agree based on these facts that fear shown during learning could be a sign of positive learning. Interested readers are referred to Appendix G, Tables 1 to 11, which provides feedback samples from the participants.

The proposed model estimated that sadness emotion was prevalent for 21.57% (equivalent to 11) of the participants, and the mapping experiment conducted mapped sadness emotion with a confused affective state. Therefore, this study suspects that there is a possibility that the lecture confused the learners. Based on previous studies and the initial mapping experiment conducted, sadness emotion is estimated as a negative learning affect. This research confirmed this through a survey that these participants were not clear with the lecture. Therefore, this study concludes that sadness emotion and confused affective states shown during learning are negative learning affect.

Four of the participants (7.84%) were estimated by the model to have predominantly shown surprise emotions. The emotion estimator and the literature mapped surprise emotion to the affective state of engagement. Therefore, surprise emotion is estimated as a positive learning affect. Through a questionnaire given to the participants, surprise emotion was verified as a positive learning affect. Anger, disgust, and surprise were not estimated for any of the participants as prevalent during the lecture, although these emotions were displayed at some points during the lecture. The summary of the analysis of the different emotions of the participants is shown in Table 5.6.

Table 5.6: Live Testing : Emotion Estimates

Emotion	No. of Participants	Percentage (%)
Fear	32	62.75
Happy	4	7.84
Sadness	11	21.57
Surprise	4	7.84
Total	51	100

Thirty-eight of the 51 participants (74.51%) were estimated to be predominantly exhibiting engaged affective states during the lecture. This shows a state of positive learning by the learner during the lecture and shows that the content is clear to the learners during that time frame of the lecture. However, 21.57% of 51 participants were also estimated to be exhibiting confusion affective state at some point during the live test. This suggests that there were time frames where the learners were potentially confused during specific times. More analysis can be performed to gather those exact times. In addition to the above, 2 of the 51 participants were estimated to be exhibiting a frustrated affective state. These results emphasise that, during

learning, the rate of assimilation is different for different learners. This may be due to external factors eg., the background of the learner. Table 5.7 provides the results of live testing.

Table 5.7: Live Testing: Affective State Estimates

Affective State	No. of Participants	Percentage(%)
Confusion	11	21.57
Engagement	38	74.51
Frustration	2	3.92
Total	51	100

This study further performed a cross-tabulation between the model's estimates of basic emotion and affective states to determine if there is an association between these variables and the following observations were reported. 81.58% of 51 participants, whose emotions were estimated as fear, were also estimated as engagement affective state. This outcome further verifies our previous mappings of fear and engagement, which is a positive learning affect. 10.53% of the participants, whose emotions were estimated as happy, also had an estimate of their affective state as engagement. Happy and engagement have been mapped as positive learning affects, and this also verifies the mappings. Sadness was associated with confusion by cross-tabulation, and 81.82% of the participants whose emotions were estimated as sadness were also estimated as confusion. This also shows that sadness and confusion are negative learning affects, as verified. Some participants (7.89%) whose affective states were estimated to be engagement were also estimated to be surprised emotions; this mapping is also verified through this live test. Live test estimates showed that there is an association between the learner's facial emotion and affective states, given a p-value of 0.000. Cross-tabulation details are tabled in Table 5.8.

5.6.3 Questionnaries: Analysis of the responses obtained

After the participants watched the video, questionnaires were provided to assess the mapping carried out by this study. The responses to the participant questionnaires are given in the following subsections, using different themes for each group of questions.

Participants' perception, learning affect and rating of the lecture

The purpose of the questions in this section is to determine how participants felt about the lecture they had watched and what learning affect they may recommend based on their observations. The questions include a suggested rating of the lecture in terms of its presentation. This study compares the learning affect indicated by the participants with the estimate made by the proposed framework, as well as confirms the various mappings. This study used Stata version 14.0 software for analysis. The responses to the questions were analysed and included in the appendix. Question

Table 5.8: Cross-tabulation between Emotion and Affective state Estimates

	Affective State Estimate			
Emotion Estimate	Confusion	Engagement	Frustration	Total
Fear	1 9.09	31 81.58	0 0.00	32 62.75
Happy	0 0.00	4 10.53	0 0.00	4 7.84
Sadness	9 81.82	0 0.00	2 100.00	11 21.57
Surprise	1 9.09	3 7.89	0 0.00	4 7.84
Total	11 100.00	38 100.00	2 100.00	51 100.00
Fisher's exact= 0.000				

1 was included to gather the participants' feelings regarding the lecture. 50.98% of the respondents were happy with the lecture, 31.37% felt neutral about the class, while 11.76% depicted a surprise emotion as their response. 5.88% of the respondents remarked they were not happy (sad) with the lecture. This study observed that a higher percentage of respondents were happy, neutral, or surprised with the lecture. These same respondents were those whose facial expression was estimated as fear emotion by the proposed framework. Through verbal questions with these participants, it was observed that participants who had fear as an estimated emotion by the proposed framework and whose responses were happy, neutral, or surprised understood the video lecture they had watched. Contrary to that, respondents who commented they were not happy with the lecture and whose facial expressions were estimated as sadness emotion by the proposed model did not understand the lecture. Details of the analysis are provided in Table 1 of Appendix G.

This study further investigated the estimates reported by the framework and the estimates that the participants identified as their feelings about the lecture they had watched. We achieved this by a cross-tabulation of the two observations. Analysis revealed that 62.75% of the participants whose emotions were estimated as fear by the model estimated their own emotions during learning as happy, neutral, or surprised. From the previous mapping of basic emotions, based on a literature survey by Kapoor et al. [26] and Zakka et al. [27], a test carried out using DAiSEE dataset, fear was considered a positive emotion. The proposed model also mapped it with affective states of engagement, which is positive learning affect according to the authors in [32, 33, 34, 35, 36].

The authors in [12, 27] considered neutral a positive emotion, and the authors in [25, 26, 12, 27] considered surprised a positive emotion. The emotion estimator

mapped surprised emotion with affective states of engagement, which is also a positive learning affect, aligning with the initial mapping. Therefore, the study would like to reiterate that surprise emotion is a positive learning affect. Fisher's exact was also calculated as part of the work using stata analysis software. This software compares the association and the statistical significance of the model's estimate with the participants' estimates [263] [264]. Fisher's exact result gave a p-value of 0.006 in this case. This shows that the predictions of the emotion estimator and what the participants have assessed as emotion throughout the learning are related, and the difference is statistically significant[268]. Table 5.9 provides the details.

Table 5.9: Cross-tabulation: Model Estimates vs Participants Estimates of Emotions

	Participants' estimation of their emotions				
Model estimation of participants' emotion	Happy	Neutral	Sadness	Surprise	**Total**
Fear	19 73.08	9 56.25	0 0.00	4 66.67	32 62.75
Happy	4 15.38	0 0.00	0 0.00	0 0.00	4 7.84
Sadness	1 3.85	6 37.50	3 100.00	1 16.67	11 21.57
Surprise	2 7.69	1 6.25	0 0.00	1 16.67	4 7.84
Total	26 100.00	16 100.00	3 100.00	6 100.00	51 100.00
Fisher's exact= 0.006					

Question 2 tries to explore the category of learning affect that the learner would like to suggest based on their estimated emotions. According to the responses, 38 (74.51%) of the 51 respondents remarked that they had a positive learning affect. These comprised respondents whose facial emotions were estimated as fear, happy, and surprise. Taking into account the mappings from the literature, and the results of the survey, the mapping of fear emotion as a positive learning affect was finalised.11 (21.57%) of the 51 respondents whose facial emotions were estimated as sadness confirm that they were not clear about the content of the lecture. In the literature, sadness emotion was considered a negative emotion. Therefore, the study validates sadness as a negative learning affect. 2 (3.92%) of the 51 respondents felt neutral about the lecture, which they also classified as their learning affect. The details of the analysis are shown in Table 2 of appendix G.

Question 3 tries to explore the aspects around the presentation of the lecture and whether the presentation influenced the learning affect. 3.92% of the respondents rated the presentation at 40%, 9.80% rated the presentation at 50%, 11.76% rated the presentation at 60%, 29.41% rated it 70%, 25.49% rated at 80%, 15.69% rated

at 90%, while 3.92% rated at 100% effective. From the observation of the ratings, it was revealed that the highest rating was between 60 and 80%, indicating a well-presented lecture based on the high response to the learning affect as positive in question 2. This emphasises the correlation between the presentation method of the lecture content and the learning affect experienced by learners. Complete results on this aspect are presented in Table 3 of appendix G.

Participants' perceptions about facial expression as an evaluation method in e-learning

The questions in this category consist of yes and no responses. The questions posed in this section are intended to verify whether facial emotion is a good method to use in the evaluation of learners in online learning and whether the method could improve the e-learning platform and possibly provide insight into the learner's learning.

From the responses to question 4, 38 participants agreed that facial emotions has the potential in evaluating learners in online learning, representing 74.51% of the total participants. 17.65% did not agree with this theory as shown in Table 4 in the appendix section. On the other hand, 43 of the 51 participants, representing 84.31% of the total responses, responded to Question 5 of the questionnaire, indicating that the learning affect of the learners could be assessed by their facial expressions. Even though only 6 respondents (3.92%) of the total participants disagreed with this notion while one participant did not give any response. However, most of the participants agreed, resulting in a significant outcome as tabled in Table 5 in Appendix G.

According to 45 respondents (88.24% of the total participants) for question 6, estimating a learner's learning affect using his or her facial emotions as a method could provide insight into a learner's comprehension level. Only three respondents (5.88%) disagreed with this statement, as shown in Table 6. The authors observed that the majority of responses were positive, indicating that most of the participants agreed that estimating the learner's learning affect using their facial emotions is a method that could provide insight into the level of comprehension of the learners.

Responses from question 7, which seeks to verify if teachers could benefit from this method in terms of learner evaluation, 41 (80.39%) of the respondents agreed with this notion, while 4 (7.89%) of the participants gave negative responses, thereby disagreeing as shown in Table 7 in Appendix G. In conclusion, the highly positive responses from this study, confirm that facial emotions are a good metric for evaluating learners during e-learning. This could also help teachers better understand their learners during online teaching sessions to a reasonable extent. Consequently, the outcome of this study is expected to generally contribute to improving e-learning.

Participants' comments regarding the proposed model

The questions in this category are designed to elicit comments from the participants on how the proposed system could be improved. The questions specifically ask for suggested criteria other than the ones explored in the proposed framework and could be used to estimate learning affect during online learning. The respondents

were also asked if there is room for improvement. The participants' opinions on the model's feedback and their own feelings were also sought with the questions.

Twenty-two (22) responses were collected for question 8. Based on the responses collected from the participants for this question, 5 respondents suggested that the learner's attention spans could be used to measure the learning affect during online sessions. This is an area that has been considered before by Rahul et al. [269, 270] and could also be improved on. Another proposal offered by 4 of the respondents was that learner engagement levels should be evaluated while learning. Our experiment found that one of the affect emotions that was correlated with happy and surprise emotions was an affective state of engagement. Some authors have also examined the level of learner engagement to determine the learning affect during the learning process. These authors include Adebi et al. [219], Mohamad et al. [271] and Bhardwaj et al. [272] among others. Three (3) of the 22 respondents suggested that verbal discussions can be held after lectures to further estimate the learning effect. They also suggested that asking learners questions in the form of a quiz could augment the estimation of their learning affect. Two (2) of the respondents stated that blinking or eye contact can be used as a modality, which is also a feasible option as shown by Sharma et al. [273]. Therefore, this study agrees that it could be developed further. According to 1 of the respondents, the detection of stress levels was another means for measuring learner's learning affect. This study would like to propose the use of multi-modal biometrics in the e-learning framework as an area of future work to contribute to the assessment of the learning affect of learners in the education domain. The details of the analysis are provided in Table 8 of Appendix G.

Based on the responses collected for question 9, 44 of the 51 participants, representing 86.2% of the total participants, agreed that the system could be improved in the future, while 6 did not know what to say and none of the participants disagreed with this notion. As a result, the researcher concluded that the system has the potential for improvement in the future. The report of this analysis is shown in Appendix G (Table 9).

The survey showed that 26 respondents (50.98%) responded to question 10. Out of these 26, 2 of the respondents gave similar responses, suggesting that an oral examination be conducted as a way by which the system could be improved. Another set of 2 respondents gave a similar response which suggested that the stress level of the learners could be detected. 2 responses did not know what to say, while none of the remaining 20 responses was the same. Some of the suggestions made by the remaining respondents (39.21%) include the body movement of learners, the mood of learners, and making observations at different times. Physical examination, sitting position, body gesticulations, the content of the lecture, and engagement with the monitor, among others, were suggested. Details of this aspect are provided in Table 10 in Appendix G.

Thirty-six (72.55%) of the total participants responded to question 11 by stating that their learning experience correlates with the feedback from the proposed model. The remaining 27.45% did not respond to the question about the relationship between their experience and what the model predicted. Based on the opinion expressed by the majority, the proposed framework is performing reasonably well

in terms of the estimations. The report of the analysis to question 11 is shown in Table 11 in Appendix G.

Cross-tabulation between observations

A cross-tabulation between observations from the survey was undertaken to determine the associations between the various observations from the survey. The work selected the association between the categorization of participants' learning affects and other variables of interest in the survey using the stata analysis tools, given earning affect experienced by the learners is the core aspect.

Association between Q1 (participant's feeling about the lecture and Q2 (learning affect categorisation)

Based on Fisher's exact test, there is a statistically significant association (p-value = 0.000) between the participant's feelings about the lecture and their learning affect. More than half of the participants with positive learning expressed happy (68.42%) and surprise (18.8%) emotions. The feeling of sadness (27.27%) was also predominant among the participants who indicated a negative learning affect during the live test. Table 5.10 shows the results of the analysis performed on cross-tabulation between questions 1 and 2. These results enabled the author to verify that emotions have a significant effect on the learning affect experienced by the learner.

Table 5.10: Cross tabulation of questions 1 and 2

	Q2			
Q1	Negative	Neutral	Positive	**Total**
Happy	0 0.00	0 0.00	26 68.42	26 50.98
Neutral	6 54.55	2 100.00	8 21.05	16 31.37
Sadness	3 27.27	0 0.00	0 0.00	3 5.88
Surprise	1 9.09	0 0.00	4 10.53	5 9.80
Total	11 100.00	2 100.00	38 100.00	51 100.00
Fisher's Exact= 0.000				

Association between Q4 (participant's perception about facial emotion expression as an evaluation method in online learning and Q2 (participant's learning affect categorisation)

The Fisher's test was performed, as an evaluation to find the association between the perception of facial expression by participants as an evaluation method in online

learning and the categorisation of learning affect as shown in Table 5.11.

Table 5.11: Cross-tabulation of questions 2 and 4

	Q2			
Q4	Negative	Neutral	Positive	**Total**
DN	0 00.00	1 50.00	3 7.89	4 7.84
No	4 36.36	0 0.00	5 13.16	9 17.65
Yes	7 63.64	1 50.00	30 78.95	38 74.51
Total	11 100.00	2 100.00	38 100.00	18 100.00
Fisher's Exact= 0.129				

A total of 78.95% of the participants who indicated a positive learning affect during the live test thought that facial emotion expression could be used as an evaluation method in online learning as opposed to 13.6% with a negative affect with similar views, while the remaining 7.89% did not know what to say. When the analysis was carried out, this study observed that the differences were not statistically significant (p-value = 0.129). This suggests that there was no significant association between the learning affect categorization and their perception of facial expression as an evaluation method in e-learning.

5.7 Discussion

Estimating a learner's learning affect, using their facial emotion and affective state estimates, learning affect estimate and feedback using both narrative reports and Q&A for teachers and learners is an important strategy in e-learning. The author of this study expects that this strategy will improve e-learning platforms that currently lack direct and appropriate feedback during online learning sessions. This study proposed a spatio-temporal reasoning framework to tackle this problem.

The proposed framework in this study could classify seven basic emotions of anger, fear, disgust, happy, neutral, sadness, and surprise from DISFA+ dataset, reporting an accuracy of 92%, an F-1 score of 88% on 2,274 sequence samples. While for the affective state classification, an accuracy of 86% and an average F1 score of 87% on 4,305 sequence samples were reported. Classified emotions and affective states were mapped further onto learning affect ie., positive, negative, and neutral categories based on a survey of relevant literature. This revealed that emotions and affective states shown by learners during learning could indicate some levels of learning impact. These emotions and affective states shown during learning will assist the teacher in either changing his or her style of teaching or building on the style to achieve better results. The outcomes will also assist the learners in realizing their weak and strong points and therefore considering ways of improvement.

This study tested the trained emotion estimator with affective state samples (DAiSEE dataset) that were not part of the training and testing to investigate the previous mappings. The testing revealed that emotions and affective states are related and are relevant to revealing the learning affect. Upon successful mapping, narrative feedback reports were generated. The input for feedback reports included a template along with estimated emotion and affective affect, mapped learning affect, and video metadata (length of the video, video labels etc.). The narrative feedback will provide information on the learning achievement of both teachers and learners. Further feedback to both teachers and learners was available through the BERT Q&A model which processed the narrative report, and a user's question to generate an answer (with reasoning embedded). The question and answer is a further way of obtaining feedback from the system. The study explored further verification of the framework through a live test and survey questionnaires administered to willing participants. Results of the study on the estimated learning affect by the system based on the questions and answers were compared with the responses obtained from the learner survey. This shows that the proposed model produced a remarkable result when tested with a real-life sample.

During the live test, 32 of the 51 participants showed fear emotions as the most frequent emotion, while 38 showed engagement as the most frequent affective state, as classified by the proposed framework. The exhibition of fear emotion connoting a positive learning affect through various experiments shows that fear shown during learning means positive learning. Engagement and fear were mapped through the various experiments as positive learning, which shows that fear and engagement are both indications that the learner understood the lesson. The survey responses enable the author to verify and validate the initial mapping between emotions/affective states and the learning affect. The work concluded with cross-tabulation between the participants' responses and the general perception of the proposed fared along with future suggestions.

5.8 Chapter Summary

The results reported from various experiments carried out to test the proposed e-learning framework were presented in this chapter. The results presented include test experiments with DISFA+ and DAiSEE datasets whose goal was to evaluate the model. These results were comparable to those reported in the literature. The result of the mapping experiment using another set of DAiSEE dataset, which aims at verifying the mapping of basic emotions to learning affects and basic emotions to the corresponding affective states, were presented. The experiment was able to show that emotions and affective states correlate and provide some information.

The chapter also discussed the integration feedback module, which includes narrative report generation and Q&A BERT-based module for individual learners and a group of learners. This was followed by live experiments to test the efficacy of the proposed framework. Live test results revealed that, when learners are learning, they exhibit some emotions and affective states that depict their comprehension of the lecture. The proposed system is expected to provide relevant feedback through the estimation of facial emotions and affective states, the mapping of estimated

emotions and affective states onto learning affect, and feedback mechanisms. Feedback throughout the assessment and learning process is intended to help learners and teachers become better learners and teachers. E-learning sessions are like blind teaching and learning sessions, in which teachers and learners do not have any physical connection to each other during learning.

This study, therefore, envisaged that the results reflected in this chapter will provide insight into understanding the level of comprehension of the learner using a diverse set of cues and feedback. Based on the participants' feedback, we believe that the proposed framework is effective and has the potential to be improved in the future, thus enhancing the e-learning platforms. Chapter 6 provides the conclusion to this study.

Chapter 6

Conclusion

6.1 Introduction

Computers and the internet in learning have brought about a significant improvement in education. Including these digital systems has been the key driver of this fundamental change in education, especially with e-learning. Although e-learning provides several benefits, including flexible learning from any geographical location, self-paced learning, and access to an interactive medium, these systems come with some pedagogical challenges. One of the significant shortcomings of e-learning platforms is the lack of reasoning with questions and answers to estimate the learner's learning affect. This study has presented a work to address this lack of reasoning. This chapter concludes the study. We divided the chapter into four sections. These sections include 6.2 which provides a summary of the work done in this study, 6.3 discusses the various observations from the experiments and results reported and 6.4 provides the conclusion of the study. Section 6.5 suggests some areas for future works.

6.2 Summary

Works that applied various machine learning algorithms to reasoning tasks were identified in this study. We considered work that used video samples as input, including relational reasoning using static images and spatio-temporal reasoning. This study also evaluated the work done in education using machine learning tools that focused on the recognition of the facial emotions of learners to determine their learning affect. However, these studies did not address the reasoning within the e-learning system.

In addressing the challenge of reasoning in e-learning systems, and achieving the objectives of this study, this study sought to combine facial emotion, affective states, learning affect estimation, and reasoning through question answers format. An approach was adopted that uses a CNN-BiLSTM that accepts a video of facial cues, and a BERT-Q&A bot that takes text input to provide the answer. These answers could reflect or estimate the learning affect of these learners. Consequently, an e-learning system that could estimate the learning affect of a learner and provide

real-time feedback was presented in this study. To reason about the learning affect of learners in an e-learning session, the system uses facial expression detection, affective state detection, mapping of estimated emotion and affective state to learning affects, and reasoning questions and answers. This was achieved using CNN-BiLSTM on the DISFA+ dataset and the DAiSEE dataset and a BERT-Q&A bot for questions and answers, with live tests to evaluate the proposed system further. The study proposed a spatio-temporal model.

Anger, disgust, fear, happy, sadness, surprise, and neutral were the seven basic emotions depicted in the test experiments carried out on the trained emotion classifier, using DISFA+ dataset and a sample size of 2,274. The classification accuracy was reported to be 91%. Reported were an average F-1 score of 88% and an average precision of 89%. For the affective state classification, boredom, confusion, engagement, and frustrated were the four affective states depicted in the test experiments conducted on the trained affective state classifier using DAISEE dataset. An accuracy of 86%, an average precision of 88% and an average F1 score of 87% were reported on 4,305 sequence samples. Based on these results, this study reiterates that the approach adopted to train the model produced a result that is comparable to those of similar studies.

Based on the mappings found in the literature, the estimated emotions and affective states were then mapped to learning affects such as positive, negative, and neutral. Mapping experiments, which map basic emotions with affective states, were further carried out, and the results show that emotions and affective states are related and present during learning.

For the feedback aspects, in the first phase, a narrative text (report) embedding reasoning was generated based on the estimates (i.e., emotion, affective state, learning affect) and video annotations (i.e., title, labels, etc.). The second phase of feedback involved using generated narrative feedback text, a question posed by a user (a teacher or a learner), and a pre-trained BERT-based language model to process and produce reasoning-based answers reflecting on a learner's understanding and/or acceptability of a lecture session.

To confirm the accuracy of the mappings and feedback mechanisms, the final framework was also tested on live samples that were collected in a lab environment. During the live test experiment performed on the proposed model, this study observed that the most frequent and predominant facial cues estimated by the model during a lecture were fear, sadness, engagement, and confusion. Based on the mapping of basic emotions with affective states, fear and sadness were mapped with affective states of confusion, which were considered learning emotions, and confusion is a negative learning affect. In contrast, this study confirmed that fear is a positive emotion based on the responses to the survey questions. This confirmation was related to the results, where 31 of the 51 participants exhibited fear emotion and an affective state of engagement. The results of the survey questions showed that these participants understood the video lecture well. Through these responses, this study could conclude that fear and engagement are positive learning affect. These results were also attributed to the literature survey findings, which highlighted that fear was considered a positive affect. Overall, 72.55% of the participants reported that their understanding of the lecture and learning-affect experience corresponded

with the model's feedback.

6.3 Observations

Consequently, the analysis of the live test showed through the estimates that there is an association between the learners' facial emotion and affective states, as shown in Table 5.8. The predictions of emotions by the proposed model and the participants' assessments of their own emotions throughout learning are related, and the analysis provided a statistically significant value as shown in Table 5.9, while Table 5.10 also revealed that emotions have a significant effect on the learning affects of learners. We, also, observed that the learning emotions/affective states shown while learners were learning during the live experiment were fear, sadness, engagement, and confusion. Could this, therefore, indicate that, when learners are learning, they mostly show these emotions and affective states? This is an area that could be further researched. Some emotions and affective states were not shown during learning, which could suggest that there are learning emotions/affective states and out-of-classroom emotions/affective states. This could also be further researched. The outcomes of the live testing demonstrated the viability of the proposed e-learning platform framework. The general results of the analyses showed that there is a correlation between the facial cues shown during learning and the learning affect and; our results show that including reasoning through Q&A to estimate the learning affect gives a deeper understanding of the learner's learning affect, especially during online learning that does not involve physical contact. Therefore, the research questions were answered. Reasoning about the learning effects of learners in an e-learning setting is therefore crucial.

6.4 Conclusion

This study establishes the foundation for a reasoning system that estimates learners' learning affects based on their facial cues and questions and answers as feedback. This demonstrates that, with the right deep learning framework, such a system is indeed possible, and its realization is imminent. It is hoped that this work will eventually lead to the development of useful tools to improve teaching and learning in e-learning, which has been difficult due to a lack of physical contact. The system is intended to produce better teachers, benefit learners, and ultimately improve e-learning. Therefore, we conclude based on the results of these experiments that applying deep learning architecture through facial emotional expression and affective state recognition, mapping estimated emotions and affective states into learning affect, and incorporating relational questions and producing answers will yield improved results for estimating the learners' comprehension state. This implies that the objectives stated have been achieved to a reasonable extent. The current work hypothesizes that reasoning can be useful in the e-learning domain and has the potential to improve existing platforms by incorporating feedback mechanisms as reasoning answers to relational questions. Therefore, we reject the null hypothesis in 1.3.1 and support the alternative hypothesis in 1.3.2.

6.5 Future Works

In future studies, modalities such as gaze at the eyes, head pose, physiological signals, body posture, and many other modalities with reasoning questions and answers could be examined to assess the learning affect of learners on an e-learning platform. Various surveys conducted in this work underpin the proposed future work. A model that could identify a learner's stress level while learning could also be developed. The learner's attention span is also an area that could be further examined. Including voice activation input to pose questions and reasoning answers to estimate the learning affect of learners is another area that could be explored.

Appendices

.1 Appendix A: Participants Information sheet

Spatio-Temporal Reasoning for Estimating Student's Learning Affect: An Approach to Strengthen E-learning.

Participant Information Sheet

Dear Sir / Madam,

My name is Christine Asaju and I am a Ph.D. student in the school of Computer Science and Applied Mathematics at the University of the Witwatersrand, Johannesburg. The title of my research is "Spatio-Temporal reasoning for Estimating a Student Learning Affect: An Approach to strengthen E-learning. As part of the studies, I am investigating the emotional changes in students during an E-learning class under the supervision of Prof. Hima Vadapalli. This study intends to improve the e-learning domain by incorporating reasoning into the existing system through the analysis of facial emotions, and questions regarding student's understanding and learning affects estimation. Estimation of learning affects means having an understanding of the attitudes, the interests and values that the learners has achieved during learning. Learning affect estimation is also a strategy adopted by educators which involves understanding a learner's emotional influence and intervening with the right method that will manage and guide the learners rightly. The study presume that by incorporating the use of facial emotion recognition, learning affect estimation and reasoning through questions answers regarding a student's understanding, it will give a better feedback to teachers, and eventually, cause an improvement on the platform. The researcher trained, validated and tested the model that was developed, by using the DISFA+ dataset.

The live testing phase of this work involves, facial emotion detection, and classification of the emotions expressions which will be acquired from live participants. The work considers mapping of the emotions that is being classified during learning into positive, negative and neutral learning affects and reasoning questions answers based on the emotions that is being detected. Positive learning affects will potentially indicate that the learners have achieved positive values from the teaching, negative learning affects will potentially indicate negative values achieved from the lecture. Neutral learning affect means neither positive or negative values was achieved from the lecture. It is expected that the proposed system will be able to provide relevant feedback by encoding text, video and reasoning, in order to interpret learning affects and provide an insight into a learner's comprehension level.

As part of this project, I would like to invite you to take part in a data collection process and answering a questionnaire. This activity will take place in the 3rd floor, Msc and PhD. computer laboratory of the school of Computer Science and applied Mathematics, T.W Kambule Mathematical Sciences building of the University of the Witwatersrand. The process involves watching a video of lecture of a topic in machine learning course and will take around 10 minutes. The content of the video will not in any way be disturbing, or emotionally manipulative and no potential triggering effects. While you watch the lecture, your facial emotion expressions will be detected by the model and the highest occurring emotion will be used to estimate the learning affects through reasoning questions answers provided by the system. With your permission, I would also video record your facial emotion changes during the data collection process using a digital device. This recording will be stored in an external hard drive and only the researchers will have access to this recording. It will be deleted at the end of the study and publications.

There will be no personal costs to you if you participate in this project, You will not receive any direct benefits from participation but there are no disadvantages or penalties if you do not choose to

Figure 1: Participants' Information Sheet

participate or if you withdraw from the study. You may withdraw at any time or not answer any question if you do not want to. The questionnaire will be completely confidential and anonymous as I will not be asking for your name or any identifying information, and the information you give to me will be held securely and not disclosed to anyone else. I will be using a pseudonym (false name) to represent your participation in my final research report. If you experience any distress or discomfort at any point in this process, we will stop the recording process or resume another time.

If you have any questions during or afterwards about this research, feel free to contact me on the details listed below. This study will be written up as a dissertation which will be available online through the university library website. If you wish to receive a summary of this report, I will be happy to send it to you (optional). With your permission the data collected from this research project may be used by other researchers in an anonymized format. If you have any concerns or complaints regarding the ethical procedures of this study, you are welcome to contact the University Human Research Ethics Committee (Non-Medical), telephone +27(0) 11 717 1408, email hrecnon-medical@wits.ac.za

Yours sincerely,
Christine Asaju

Researcher: Christine Bukola Asaju, 1990591@students.wits.ac.za, 0632064007

Supervisor: Prof. Hima Vadapalli, hima.vadapalli@wits.ac.za, 0117176186

Figure 2: Participants' Information Sheet (continued)

.2 Appendix B: Consent Form

Consent Form

Title of project**: Spatio-Temporal Reasoning for Estimating Student's learning outcome: An Approach to strengthen E-learning**

Name of researcher **: Christine Bukola Asaju**

I, ……., agree to participate in this research project. The research has been explained to me and I understand what my participation will involve. I agree to the following:

(Please circle the relevant options below).

I agree that my participation will remain anonymous	YES	NO
I agree that the researcher may use anonymous quotes in his / her research report	YES	NO
I agree that my participation in the data collection process may be video recorded	YES	NO

…..……(signature).....(name of participant)
…… (date)
…… (signature).....(name of person seeking consent).....(date)

Figure 3: Consent form

.3 Appendix C: Questionnaire

Spatio-Temporal Reasoning for Estimating Student's Learning Affect: An Approach to Strengthen E-learning.

Questionnaire

This study intends to improve upon the current set-up in the e-learning environment by the analysis of facial emotion recognition, learning affect estimations and reasoning through questions answers regarding a student's understanding. The title of this research is "Spatio-Temporal reasoning for Estimating a Student Learning Affect: An Approach to strengthen E-learning. As part of the studies, I am investigating the emotional changes in students during an E-learning class under the supervision of Prof. Hima Vadapalli. This study proposed to improve the e-learning domain by incorporating reasoning into the existing system through the analysis of facial emotions, and questions regarding student's understanding and learning affects estimation. Estimation of learning affects means having an understanding of the attitudes, the interests and values that the learners has achieved during learning. Learning affect estimation is also a strategy adopted by educators which involves understanding a learner's emotional influence and intervening with the right method that will manage and guide the learners rightly. The work developed a spatio-temporal model to handle video and text input.

In the study after conducting a live test, which includes detection and classification of the emotions,will map the emotions that is being expressed into positive, negative, and neutral learning affects. The intention is to apply this system to the e-learning domain in order to estimate the learning affects of students.It is presumed that the incorporation of facial emotion recognition, learning affect estimation and reasoning through questions answers will improve the current set up in the e-learning domain.

As part of the work, the researcher would like you to answer the following questions in attempt to carry out a survey to get a further understanding of the accuracy of the system that was trained, validated, tested and which a live testing is being conducted.

Please answer the following questions which is based on the tutoring video, the emotion and the learning affect estimation. Please note that your answers will be kept anonymous.
Thank you.

Questions	**Answers**							
1. How do you feel about the lecture? Please respond to this question by picking from the list of emotions given.	Anger	Disgust	Fear	Happy	Neutral	Sadness	Surprise	
2. What category of learning affect will you classify your learning outcomes from the lecture?	Positive	Negative	Neutral					
3. Give a rating of the lecture you have watched. For example, how well it was presented.	30%	40%	50%	60%	70%	80%	90%	100%
4. Is facial emotion expression a good method to evaluate a student in an online learning?	YES	NO	I don't know	Comment				
5. Do you think it is a good measure to improve upon the e-learning environment by estimating the learning affects of students through their facial emotion expression?	YES	NO	I don't know	Comment				
6. Can this method provide an insight to learner's comprehension level.	YES	NO	I don't know	Comment				

Figure 4: Questionnaire page 1

7. Can this study help teachers to adequately evaluate the students during online teaching sessions?	YES	NO	I don't know	Comment
8.Suggest other criteria that could be used to estimate learning affect in online learning platform.				
9. Can this system be improved upon in the future?	YES	NO	I don't Know	Comments
10. If yes for question(9)Suggest ways by which it can be improved upon.	It can be improved upon by…… ……..			
11. Do you think that your learning experience is inline with the model's predictions?	YES			NO

Figure 5: Questionnaire page 2

.4 Appendix D: Approval Letter from head of School

School of Computer Science and Applied Mathematics

To
The Ethics Committee
Wits University

14 July 2021

Dear committee members,

Letter of Support for Ms Christine Asaju

I, Prof. Hima Vadapalli, Acting Head of CSAM, would like to take this opportunity to support Ms. Asaju's application requesting for ethics clearance for her project titled "Spatio-Temporal Reasoning for Estimating Student's Learning Affects: An Approach to Strengthen E-learning ", clearance number: H21/05/02.

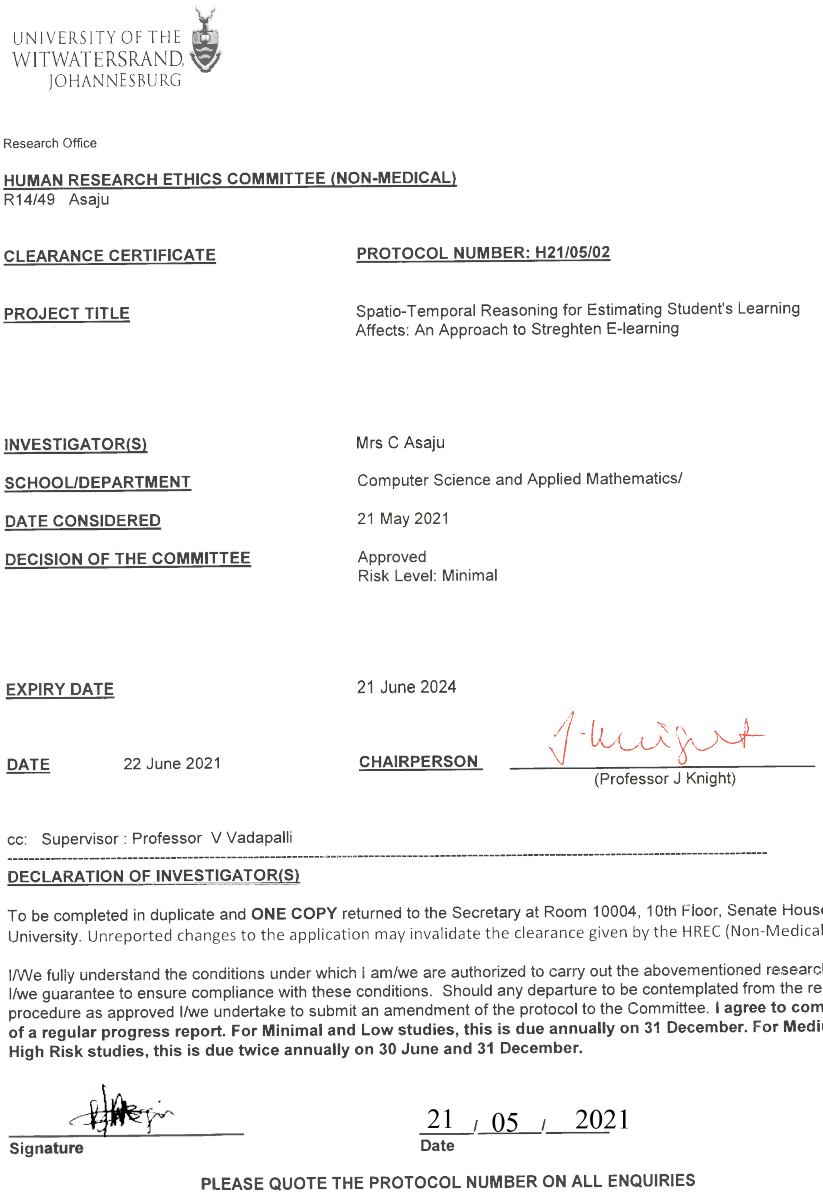
I am hopeful that allowing Ms. Asaju to collect data required for her project will lead to the fulfilment of some of the objectives of her doctoral studies and publication of results thereafter.

Kind regards,

Prof. Hima B. Vadapalli,
Associate Professor and Deputy Head of School
School of Computer Science and Applied Mathematics,
Room UG 13, Upper Ground Floor MSB, West Campus,
University of the Witwatersrand
Tel: +27(0) 11 717 6188.

Figure 6: Letter of approval from the head of school

.5 Appendix E: Ethics clearance



UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Research Office

HUMAN RESEARCH ETHICS COMMITTEE (NON-MEDICAL)
R14/49 Asaju

CLEARANCE CERTIFICATE	**PROTOCOL NUMBER: H21/05/02**
PROJECT TITLE	Spatio-Temporal Reasoning for Estimating Student's Learning Affects: An Approach to Streghten E-learning
INVESTIGATOR(S)	Mrs C Asaju
SCHOOL/DEPARTMENT	Computer Science and Applied Mathematics/
DATE CONSIDERED	21 May 2021
DECISION OF THE COMMITTEE	Approved Risk Level: Minimal
EXPIRY DATE	21 June 2024
DATE 22 June 2021	**CHAIRPERSON** (Professor J Knight)

cc: Supervisor : Professor V Vadapalli

DECLARATION OF INVESTIGATOR(S)

To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10004, 10th Floor, Senate House, University. Unreported changes to the application may invalidate the clearance given by the HREC (Non-Medical)

I/We fully understand the conditions under which I am/we are authorized to carry out the abovementioned research and I/we guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I/we undertake to submit an amendment of the protocol to the Committee. **I agree to completion of a regular progress report. For Minimal and Low studies, this is due annually on 31 December. For Medium and High Risk studies, this is due twice annually on 30 June and 31 December.**

Signature

21 / 05 / 2021
Date

PLEASE QUOTE THE PROTOCOL NUMBER ON ALL ENQUIRIES

Figure 7: Ethics clearance certificate

.6 Appendix F: Result output interface

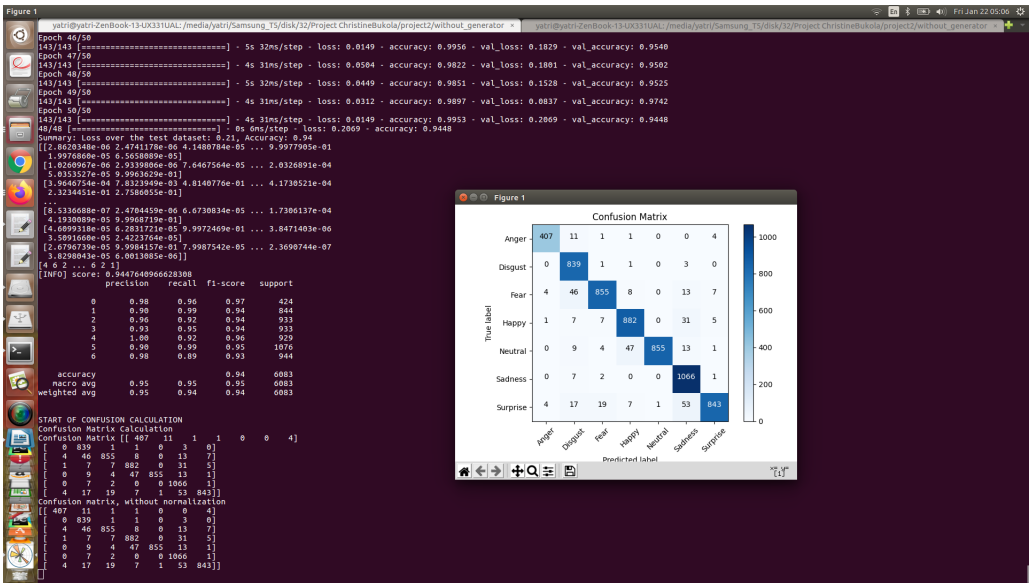


Figure 8: Training/validation result output interface for emotion sample

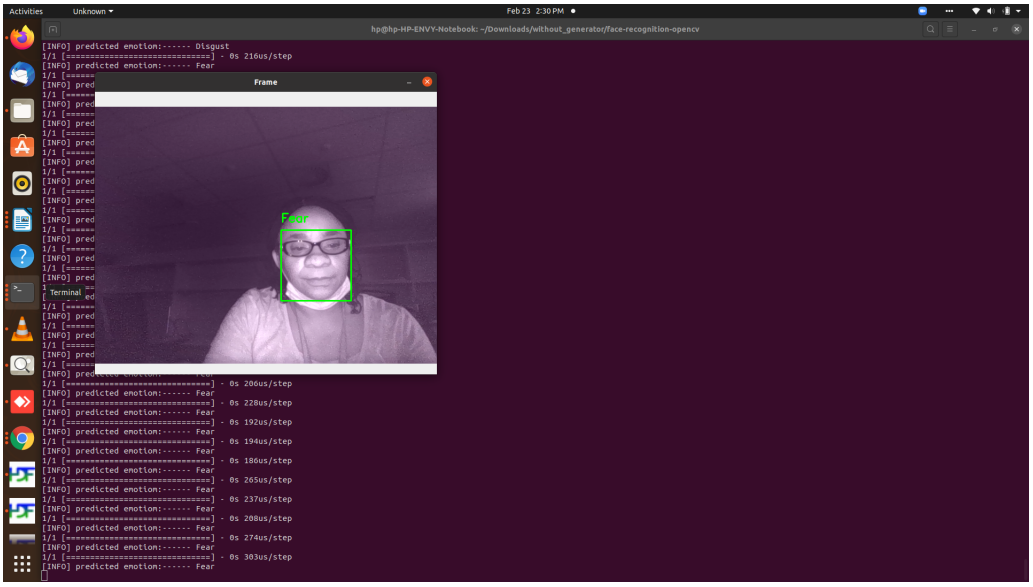


Figure 9: Emotion estimation output interface

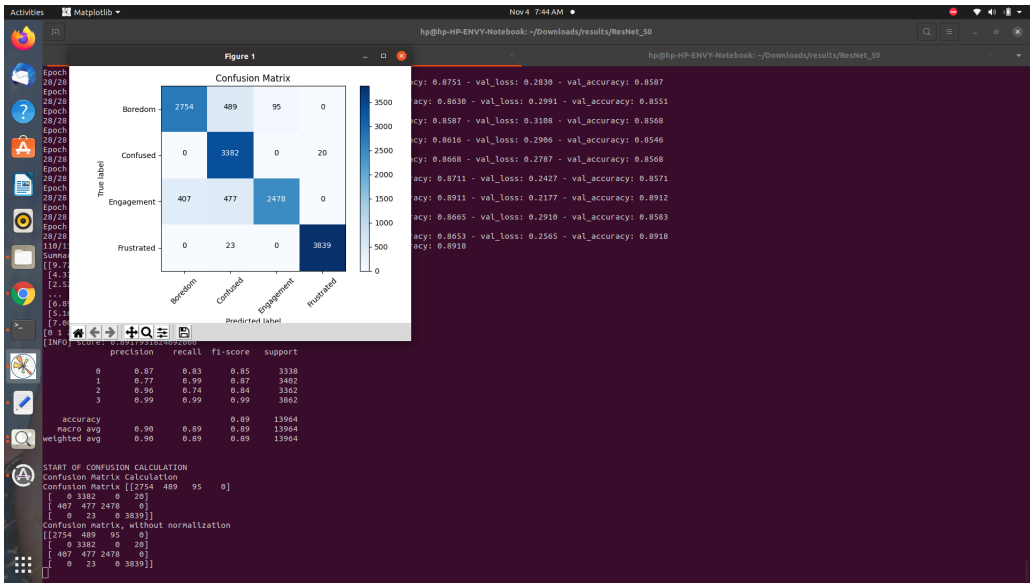


Figure 10: Training/validation result output interface for affective state sample

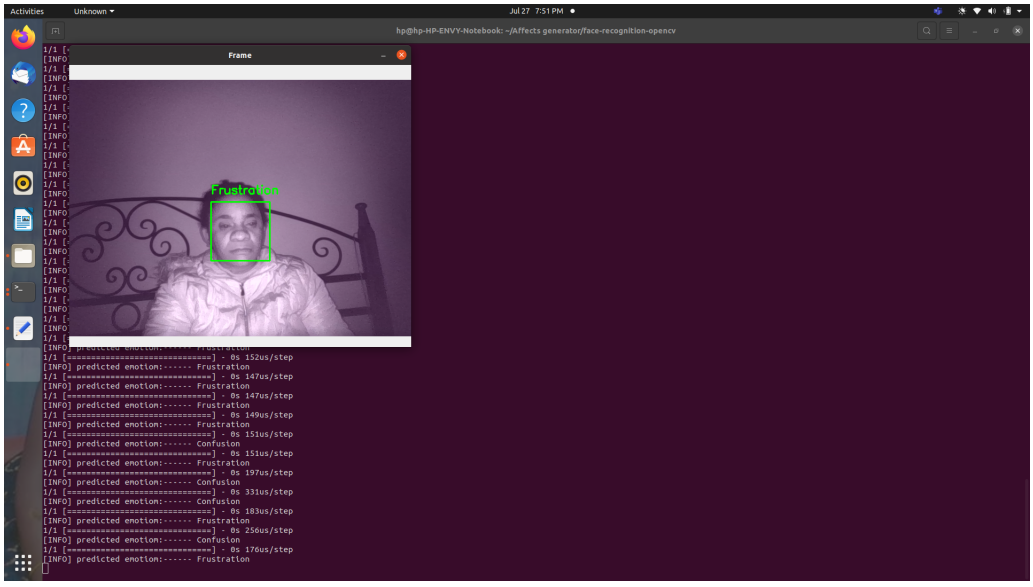


Figure 11: Affective state estimation output interface

.7 Appendix G: Analysis of Paticipants' Responses to questionnaire

Table 1: Question 1: How do you feel about the lecture?

Analysis of responses		
Responses	**Frequency**	**Percentage(%)**
Happy	26	50.98
Neutral	16	31.37
Sadness	3	5.88
Surprise	6	11.76
Total	51	100.00

Table 2: Question 2: What category of learning affect will you classify your learning outcomes from the lecture?

Category of learning affects		
Responses	**Frequency**	**Percentage** (%)
Negative	11	21.57
Neutral	2	3.92
Positive	38	74.51
Total	51	100.00

Table 3: Question 3: Give a rating of the lecture you have watched

Rating of lecture presentation		
Rating(%)	**Frequency**	**Percentage (%)**
40	2	3.92
50	5	9.80
60	6	11.76
70	15	29.41
80	13	25.49
90	8	15.69
100	2	3.92
Total	**51**	**100**

Table 4: Question 4: Is facial emotion expression a good method to evaluate a learner in online learning?

Participants' opinion regarding the use of facial emotion to evaluate students' learning affect in online learning		
Response	**Frequency**	**Percentage (%)**
No response	4	7.84
No	9	17.65
Yes	38	74.51
Total	**51**	**100**

Table 5: Question 5: Do you think it is a good measure to improve upon the e-learning environment by estimating the learning affect of learners through their facial emotional expressions?

Participants' opinion regarding the use of students' facial emotion estimation to improve the e-learning domain		
Response(%)	**Frequency**	**Percentage (%)**
No	6	3.92
Yes	43	84.31
Total	**49**	**100**

Table 6: Question 6: Can this method provide insight to learner's comprehension level?

Analysis of participants' responses		
Response(%)	**Frequency**	**Percentage (%)**
No	3	5.88
Yes	45	88.24
Total	**48**	**100**

Table 7: Question 7: Can this study help teachers evaluate students adequately during online teaching sessions?

Analysis of participants' responses		
Response(%)	**Frequency**	**Percentage (%)**
No	4	7.84
Yes	41	80.39
Total	**45**	**100**

Table 8: Question 8: Suggest other criteria that could be used to estimate the learning affect in an online platform.

Participants' responses		
Suggestions	**Frequency**	**Percentage (%)**
Verbal discussions after the lecture	3	13.63
stress level detection	1	4.54%
Blinking of the eye/eye contact	2	9.09%
Attention span	5	22.72%
Questions and answers	3	13.63%
Engagement level	4	18.18%
Don't know	4	18.18%
Total =22		100

Table 9: Question 9: Can this system be improved in the future?

Participants' responses		
Response(%)	**Frequency**	**Percentage (%)**
Yes	44	86.2
Total	**44**	**100**

Table 10: Question 10: Suggest ways in which the system could be improved.

Responses	**Frequency**	**Percentage**
Body movement	1	3.85
Mood of student	1	3.85
Movement detection	1	3.85
Observing at different times	1	3.85
Oral examination	2	7.69
Physical observation	1	3.85
Question answering	1	3.85
Sitting position	1	3.85
To focus on cummulative emotion	1	3.85
Audio features	1	3.85
Body gesticulations	1	3.85
Content of the lecture	1	3.85
don't know	2	7.69
Engagement with the monitor	1	3.85
Engagement level	1	3.85
Include questions	1	3.85
More facial features	1	3.85
Physical class	1	3.85
Physical observations	1	3.85
Reactions to lecture	1	3.85
Stress level	2	7.69
don't know	1	3.85
student's mood	1	3.85
Total	**26**	**100.00**

Table 11: Question 11: Do you think that your learning experience is in line with the model predictions?

Participants' responses		
Response(%)	**Frequency**	**Percentage (%)**
No response	14	27.45
Yes	37	72.55
Total	**44**	**100**

Bibliography

[1] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," in *Advances in neural information processing systems*, 2017, pp. 4967–4976.

[2] D. Raposo, A. Santoro, D. Barrett, R. Pascanu, T. Lillicrap, and P. Battaglia, "Discovering objects and their relations from entangled scene representations," *arXiv preprint arXiv:1702.05068*, 2017.

[3] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3076–3086.

[4] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 3942–3951.

[5] R. Krishna, I. Chami, M. Bernstein, and L. Fei-Fei, "Referring relationships," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6867–6876.

[6] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Inferring and executing programs for visual reasoning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2989–2998.

[7] B. Bennett, A. G. Cohn, F. Wolter, and M. Zakharyaschev, "Multi-dimensional modal logic as a framework for spatio-temporal reasoning," *Applied Intelligence*, vol. 17, no. 3, pp. 239–251, 2002.

[8] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, "TGIF-QA: Toward spatio-temporal reasoning in visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2758–2766.

[9] X. Song, Y. Shi, X. Chen, and Y. Han, "Explore multi-step reasoning in video question answering," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 239–247.

[10] J. Lei, L. Yu, M. Bansal, and T. L. Berg, "TVQA: Localized, compositional video question answering," *arXiv preprint arXiv:1809.01696*, 2018.

[11] O. Nafea, W. Abdul, G. Muhammad, and M. Alsulaiman, "Sensor-based human activity recognition with spatio-temporal deep learning," *Sensors*, vol. 21, no. 6, p. 2141, 2021.

[12] M. Pan, J. Wang, and Z. Luo, "Modelling study on learning affects for classroom teaching/learning auto-evaluation," *Science Journal of Education*, vol. 6, no. 3, p. 81, 2018.

[13] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *International Journal of Computer Vision*, vol. 127, no. 2, pp. 115–142, 2019.

[14] L. Krithika, "Student emotion recognition system (sers) for e-learning improvement based on learner concentration metric," *Procedia Computer Science*, vol. 85, pp. 767–776, 2016.

[15] T. Nishino, R. Ozaki, Y. Momoki, T. Taniguchi, R. Kano, N. Nakano, Y. Tagawa, M. Taniguchi, T. Ohkuma, and K. Nakamura, "Reinforcement learning with imbalanced dataset for data-to-text medical report generation," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2223–2236.

[16] J. Sun, D. Wei, L. Wang, and Y. Zheng, "Lesion guided explainable few weak-shot medical report generation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 615–625.

[17] B. Kort, "Cognition, affect, and learning: The role of emotions in learning," IEEE, pp. https://sites.google.com/site/barrykort/home/cognition–affect–and–learning, (accessed 2022).

[18] M. Mavadati, P. Sanger, and M. H. Mahoor, "Extended disfa dataset: Investigating posed and spontaneous facial expressions," in *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 1–8.

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

[20] S. Kuzdeba, J. Robinson, and J. Carmack, "Transfer learning with radio frequency signals," in *2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 2021, pp. 1–9.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[22] P. Srivastava, "Essentials of deep learning: Introduction to Long Short Term Memory," *Analytics Vidhya*, vol. 23, 2017.

[23] Z. Cui, R. Ke, Z. Pu, and Y. Wang, "Deep bidirectional and unidirectional lstm recurrent neural network for network-wide traffic speed prediction," *arXiv preprint arXiv:1801.02143*, 2018.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[25] M. Sathik and S. G. Jonathan, "Effect of facial expressions on student's comprehension recognition in virtual educational environments," *SpringerPlus*, vol. 2, no. 1, p. 455, 2013.

[26] A. Kapoor, S. Mota, R. W. Picard *et al.*, "Towards a learning companion that recognizes affect," in *AAAI Fall symposium*, vol. 543, 2001, pp. 2–4.

[27] B. E. Zakka and H. Vadapalli, "Estimating student learning affect using facial emotions," in *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*. IEEE, 2020, pp. 1–6.

[28] E. C. Daschmann, T. Goetz, and R. H. Stupnisky, "Testing the predictors of boredom at school: Development and validation of the precursors to boredom scales," *British Journal of Educational Psychology*, vol. 81, no. 3, pp. 421–440, 2011.

[29] S. D'Mello, "A selective meta-analysis on the relative incidence of discrete affective states during learning with technology." *Journal of Educational Psychology*, vol. 105, no. 4, p. 1082, 2013.

[30] S. K. D'Mello and A. C. Graesser, "Confusion," in *International handbook of emotions in education*. Routledge, 2014, pp. 299–320.

[31] N. L. Stein and L. J. Levine, "Making sense out of emotion: The representation and use of goal-structured knowledge," in *Psychological and biological approaches to emotion*. Psychology Press, 2013, pp. 63–92.

[32] J. Nakamura and M. Csikszentmihalyi, "The concept of flow," in *Flow and the foundations of positive psychology*. Springer, 2014, pp. 239–263.

[33] M. Csikszentmihalyi, "The contribution of flow to positive psychology," 2000.

[34] M. Csikszentmihalyi and R. Larson, *Flow and the foundations of positive psychology*. Springer, 2014, vol. 10.

[35] D. J. Shernoff, M. Csikszentmihalyi, B. Schneider, and E. S. Shernoff, "Student engagement in high school classrooms from the perspective of flow theory," in *Applications of flow in human development and education*. Springer, 2014, pp. 475–494.

[36] Y. Guo, B. Klein, Y. Ro, and D. Rossin, "The impact of flow on learning outcomes in a graduate-level information management course," *Journal of Global Business Issues*, vol. 1, no. 2, pp. 31–39, 2007.

[37] A. Indrasari, N. Nadjmie, and E. Endri, "Determinants of satisfaction and loyalty of e-banking users during the covid-19 pandemic," *International Journal of Data and Network Science*, vol. 6, no. 2, pp. 497–508, 2022.

[38] R. Oliver and J. Herrington, "Teaching and learning online: A beginner's guide to e-learning and e-teaching in higher education," 2001.

[39] R. Subudhi and D. Palai, "Impact of internet use during covid lockdown," *Horizon J. Hum. & Soc. Sci*, vol. 2, pp. 59–66, 2020.

[40] C. B. Asaju and H. Vadapalli, "Affects analysis: A temporal approach to estimate students' learning," in *2021 3rd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, 2021, pp. 1–7.

[41] E. Mushtaha, S. A. Dabous, I. Alsyouf, A. Ahmed, and N. R. Abdraboh, "The challenges and opportunities of online learning and teaching at engineering and theoretical colleges during the pandemic," *Ain Shams Engineering Journal*, vol. 13, no. 6, p. 101770, 2022.

[42] R. A. Brooks, "Intelligence without reason," in *The artificial life route to artificial intelligence*. Routledge, 2018, pp. 25–81.

[43] G. P. Goodwin and P. Johnson-Laird, "Reasoning about relations." *Psychological review*, vol. 112, no. 2, p. 468, 2005.

[44] P. A. Alexander, "Relational thinking and relational reasoning: harnessing the power of patterning," *NPJ science of learning*, vol. 1, p. 16004, 2016.

[45] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 803–818.

[46] G. Singh, "Spatio-temporal relational reasoning for video question answering," Ph.D. dissertation, University of British Columbia, 2019.

[47] S. Wang, J. Cao, and P. Yu, "Deep learning for spatio-temporal data mining: A survey," *IEEE transactions on knowledge and data engineering*, 2020.

[48] Y.-H. H. Tsai, S. Divvala, L.-P. Morency, R. Salakhutdinov, and A. Farhadi, "Video relationship reasoning using gated spatio-temporal energy graph," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 424–10 433.

[49] B. Akkoyunlu and M. Y. Soylu, "A study on students' views on blended learning environment," *Turkish Online Journal of Distance Education*, vol. 7, no. 3, pp. 43–56, 2006.

[50] A. A. Pise, H. Vadapalli, and I. Sanders, "Estimation of learning affects experienced by learners: an approach using relational reasoning and adaptive mapping," *Wireless Communications and Mobile Computing*, vol. 2022, 2022.

[51] D. D. Dixson and F. C. Worrell, "Formative and summative assessment in the classroom," *Theory into practice*, vol. 55, no. 2, pp. 153–159, 2016.

[52] R. Klein and T. Celik, "The wits intelligent teaching system: Detecting student engagement during lectures using convolutional neural networks," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 2856–2860.

[53] S. Scotti, M. Mauri, R. Barbieri, B. Jawad, S. Cerutti, L. Mainardi, E. N. Brown, and M. A. Villamira, "Automatic quantitative evaluation of emotions in e-learning applications," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2006, pp. 1359–1362.

[54] M. A. A. Dewan, M. Murshed, and F. Lin, "Engagement detection in online learning: a review," *Smart Learning Environments*, vol. 6, no. 1, pp. 1–20, 2019.

[55] F. H. Leong, "Deep learning of facial embeddings and facial landmark points for the detection of academic emotions," in *Proceedings of the 5th International Conference on Information and Education Innovations*, 2020, pp. 111–116.

[56] M. Magdin, M. Turcani, and L. Hudec, "Evaluating the emotional state of a user using a webcam," 2016.

[57] R. E. Ferdig, E. Baumgartner, R. Hartshorne, R. Kaplan-Rakowski, and C. Mouza, *Teaching, technology, and teacher education during the COVID-19 pandemic: Stories from the field*. Association for the Advancement of Computing in Education Waynesville, NC, 2020.

[58] S. D'Mello, B. Lehman, J. Sullins, R. Daigle, R. Combs, K. Vogt, L. Perkins, and A. Graesser, "A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning," in *International conference on intelligent tutoring systems*. Springer, 2010, pp. 245–254.

[59] L.-f. Zhang, M. Fu, D. T. Li, and Y. He, "Emotions and teaching styles among academics: the mediating role of research and teaching efficacy," *Educational Psychology*, vol. 39, no. 3, pp. 370–394, 2019.

[60] M. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, "Student engagement predictions in an e-learning system and their impact on student course assessment scores," *Computational intelligence and neuroscience*, vol. 2018, 2018.

[61] P. DeLellis, Teresa, P. M. P. Noureldin, Marwa, P. M. Park, Sharon K, P. Shields, Kelly M., P. M. Bryant, Alicia, P. P. Chen, Aleda M H, and M. P. Petrelli, Heather M W, "A situational judgment test to assess students' achievement of accreditation council for pharmacy education standards 3 and 4," *American Journal of Pharmaceutical Education*, vol. 86, no. 2, pp. 124–130, 2022.

[62] R. N. Landis and A. L. Reschly, "Reexamining gifted underachievement and dropout through the lens of student engagement," *Journal for the Education of the Gifted*, vol. 36, no. 2, pp. 220–249, 2013.

[63] S. Saha, "A comprehensive guide to convolutional neural networks – the ELI5 way," 2018.

[64] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[65] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2901–2910.

[66] A. Antoniou, A. Słowik, E. J. Crowley, and A. Storkey, "Dilated densenets for relational reasoning," *arXiv preprint arXiv:1811.00410*, 2018.

[67] R. Palm, U. Paquet, and O. Winther, "Recurrent relational networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 3368–3378.

[68] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, "Learning to reason: End-to-end module networks for visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 804–813.

[69] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 39–48.

[70] G. R. VandenBos, *APA dictionary of psychology.* American Psychological Association, 2007.

[71] Z. Chen, B. Ramachandra, T. Wu, and R. R. Vatsavai, "Relational long short-term memory for video action recognition," *arXiv preprint arXiv:1811.07059*, 2018.

[72] F. Baradel, N. Neverova, C. Wolf, J. Mille, and G. Mori, "Object level visual reasoning in videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 105–121.

[73] L. C. Pickup, Z. Pan, D. Wei, Y. Shih, C. Zhang, A. Zisserman, B. Scholkopf, and W. T. Freeman, "Seeing the arrow of time," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2035–2042.

[74] N. Cauli and D. Reforgiato Recupero, "Survey on videos data augmentation for deep learning models," *Future Internet*, vol. 14, no. 3, p. 93, 2022.

[75] A. Jelfs and C. Colbourn, "Virtual seminars and their impact on the role of the teaching staff," *Computers & Education*, vol. 38, no. 1-3, pp. 127–136, 2002.

[76] M. Jones, "Applications of artificial intelligence within education," *Computers & mathematics with applications*, vol. 11, no. 5, pp. 517–526, 1985.

[77] R. Pekrun and L. Linnenbrink-Garcia, "Academic emotions and student engagement," in *Handbook of research on student engagement.* Springer, 2012, pp. 259–282.

[78] R. Pekrun, "Academic emotions," *Handbook of motivation at school*, vol. 2, pp. 120–144, 2016.

[79] K. R. Harris, S. E. Graham, T. E. Urdan, S. E. Graham, J. M. Royer, and M. E. Zeidner, *APA educational psychology handbook, Vol 2: Individual differences and cultural and contextual factors.* American Psychological Association, 2012.

[80] A. C. Graesser and S. D'Mello, "Emotions during the learning of difficult material," in *Psychology of learning and motivation.* Elsevier, 2012, vol. 57, pp. 183–225.

[81] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 42–55, 2011.

[82] P. T. Ewell, "Accreditation and student learning outcomes: A proposed point of departure. chea occasional paper." 2001.

[83] J. Friedlander and A. M. Serban, "Meeting the challenges of assessing student learning outcomes," *New Directions for Community Colleges*, vol. 2004, no. 126, pp. 101–109, 2004.

[84] D. Hume, "Emotions and moods," *Organizational behavior*, no. 258-297, 2012.

[85] D. B. McLeod and V. M. Adams, *Affect and mathematical problem solving: A new perspective.* Springer Science & Business Media, 2012.

[86] A. Ray and A. Chakrabarti, "Design and implementation of technology enabled affective learning using fusion of bio-physical and facial expression," *Journal of Educational Technology & Society*, vol. 19, no. 4, pp. 112–125, 2016.

[87] W. J. Popham, "Assessing student affect." *Educational Leadership*, vol. 66, no. 8, pp. 85–86, 2009.

[88] ——, "Assessment literacy for teachers: Faddish or fundamental?" *Theory into practice*, vol. 48, no. 1, pp. 4–11, 2009.

[89] N. G. Lederman, "Teachers' understanding of the nature of science and classroom practice: Factors that facilitate or impede the relationship," *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, vol. 36, no. 8, pp. 916–929, 1999.

[90] J. A. DeFalco, J. P. Rowe, L. Paquette, V. Georgoulas-Sherry, K. Brawner, B. W. Mott, R. S. Baker, and J. C. Lester, "Detecting and addressing frustration in a serious game for military training," *International Journal of Artificial Intelligence in Education*, vol. 28, no. 2, pp. 152–193, 2018.

[91] R. A. Calvo and S. D'Mello, "Frontiers of affect-aware learning technologies," *IEEE Intelligent Systems*, vol. 27, no. 6, pp. 86–89, 2012.

[92] M. Ferreira, B. Martinsone, and S. Talić, "Promoting sustainable social emotional learning at school through relationship-centered learning environment, teaching methods and formative assessment," *Journal of Teacher Education for Sustainability*, vol. 22, no. 1, pp. 21–36, 2020.

[93] J. Liu, G. Wu, Y. Luo, S. Qiu, S. Yang, W. Li, and Y. Bi, "Eeg-based emotion classification using a deep neural network and sparse autoencoder," *Frontiers in Systems Neuroscience*, vol. 14, p. 43, 2020.

[94] Y. Jiang, W. Li, M. S. Hossain, M. Chen, A. Alelaiwi, and M. Al-Hammadi, "A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition," *Information Fusion*, vol. 53, pp. 209–221, 2020.

[95] R. C. Dhingra and S. Ram Avtar Jaswal, "Emotion recognition based on eeg using deap dataset," *European Journal of Molecular & Clinical Medicine*, vol. 8, no. 3, pp. 3509–3517, 2021.

[96] P. Ekman and D. Keltner, "Universal facial expressions of emotion," *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture*, pp. 27–46, 1997.

[97] D. A. Sauter, "The nonverbal communication of positive emotions: An emotion family approach," *Emotion Review*, vol. 9, no. 3, pp. 222–234, 2017.

[98] M. B. MacMillan, H. R. Field, I. Neath, and A. M. Surprenant, "Valence does not affect recognition." *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, vol. 76, no. 2, p. 111, 2022.

[99] G. K. Verma and U. S. Tiwary, "Affect representation and recognition in 3d continuous valence–arousal–dominance space," *Multimedia Tools and Applications*, vol. 76, no. 2, pp. 2159–2183, 2017.

[100] J. De Rivera and C. Grinkis, "Emotions as social relationships," *Motivation and emotion*, vol. 10, no. 4, pp. 351–369, 1986.

[101] T. T. Hasan and A. H. Issa, "Investigation of facial aggression detection techniques based on facial action units using machine learning," *Applied Nanoscience*, pp. 1–19, 2022.

[102] P. Ekman, W. V. Freisen, and S. Ancoli, "Facial signs of emotional experience." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1125, 1980.

[103] S. Katsigiannis and N. Ramzan, "Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp. 98–107, 2017.

[104] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

[105] A. John, M. Abhishek, A. S. Ajayan, S. Sanoop, and V. R. Kumar, "Real-time facial emotion recognition system with improved preprocessing and feature extraction," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*. IEEE, 2020, pp. 1328–1333.

[106] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[107] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[108] S. Akay and N. Arica, "Stacking multiple cues for facial action unit detection," *The Visual Computer*, pp. 1–16, 2021.

[109] D. McDuff, R. Kaliouby, T. Senechal, M. Amr, J. Cohn, and R. Picard, "Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 881–888.

[110] D. McDuff, M. Amr, and R. El Kaliouby, "Am-fed+: an extended dataset of naturalistic facial expressions collected in everyday settings," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 7–17, 2018.

[111] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International conference on neural information processing*. Springer, 2013, pp. 117–124.

[112] A. Kaehler and G. Bradski, *Learning OpenCV 3: computer vision in C++ with the OpenCV library*. " O'Reilly Media, Inc.", 2016.

[113] M. D. Zbancioc and S. M. Feraru, "A study about the automatic recognition of the anxiety emotional state using emo-db," in *2015 E-Health and Bioengineering Conference (EHB)*. IEEE, 2015, pp. 1–4.

[114] A. Schmitt, S. Ultes, and W. Minker, "A parameterized and annotated spoken dialog corpus of the cmu let's go bus information system." in *LREC*, 2012, pp. 3369–3373.

[115] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," *Speech Communication*, vol. 53, no. 1, pp. 36–50, 2011.

[116] N. Hossain, J. Krumm, M. Gamon, and H. Kautz, "Semeval-2020 task 7: Assessing humor in edited news headlines," *arXiv preprint arXiv:2008.00304*, 2020.

[117] S. Buechel and U. Hahn, "Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, pp. 578–585.

[118] S. M. Mohammad and F. Bravo-Marquez, "Wassa-2017 shared task on emotion intensity," *arXiv preprint arXiv:1708.03700*, 2017.

[119] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," *arXiv preprint arXiv:1710.03957*, 2017.

[120] R. Walk and K. Walters, "Perception of the smile and other emotions of the body and face at different distances," *Bulletin of the Psychonomic Society*, vol. 26, no. 6, 1988.

[121] Z. Yang, A. Kay, Y. Li, W. Cross, and J. Luo, "Pose-based body language recognition for emotion and psychiatric symptom interpretation," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 294–301.

[122] C. Savitha, U. Ujwal, and M. Smitha, "Detection of single and multi-character tulu text blocks," in *2021 IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC)*. IEEE, 2021, pp. 1–6.

[123] Y. Sun, W. Jiang, J. Yang, and W. Li, "Sar target recognition using cgan-based sar-to-optical image translation," *Remote Sensing*, vol. 14, no. 8, p. 1793, 2022.

[124] S. Saifullah and A. P. Suryotomo, "Identification of chicken egg fertility using svm classifier based on first-order statistical feature extraction," *arXiv preprint arXiv:2201.04063*, 2022.

[125] D. Yang, A. Alsadoon, P. Prasad, A. Singh, and A. Elchouemi, "An emotion recognition model based on facial recognition in virtual learning environment," *Procedia Computer Science*, vol. 125, pp. 2–10, 2018.

[126] T. Kalsum, S. M. Anwar, M. Majid, B. Khan, and S. M. Ali, "Emotion recognition from facial expressions using hybrid feature descriptors," *IET Image Processing*, vol. 12, no. 6, pp. 1004–1012, 2018.

[127] J. Guo, S. Zhou, J. Wu, J. Wan, X. Zhu, Z. Lei, and S. Z. Li, "Multi-modality network with visual and geometrical information for micro emotion recognition," in *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*. IEEE, 2017, pp. 814–819.

[128] D. V. Sang, N. Van Dat *et al.*, "Facial expression recognition using deep convolutional neural networks," in *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 2017, pp. 130–135.

[129] P. M. Ferreira, F. Marques, J. S. Cardoso, and A. Rebelo, "Physiological inspired deep neural networks for emotion recognition," *IEEE Access*, vol. 6, pp. 53 930–53 943, 2018.

[130] R. Albasrawi, F. F. Fadhil, and M. T. Ghazal, "Driver drowsiness monitoring system based on facial landmark detection with convolutional neural network for prediction," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 5, pp. 2637–2644, 2022.

[131] C. Qi, M. Li, Q. Wang, H. Zhang, J. Xing, Z. Gao, and H. Zhang, "Facial expressions recognition based on cognition and mapped binary patterns," *IEEE Access*, vol. 6, pp. 18 795–18 803, 2018.

[132] M. F. Hirzi, S. Efendi, and R. W. Sembiring, "Literature study of face recognition using the viola-jones algorithm," in *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*. IEEE, 2021, pp. 1–6.

[133] N. Arun, S. Joshua, E. Daniel *et al.*, "Facial attribute analysis for emotion recognition using deep learning," in *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2022, pp. 1435–1441.

[134] N. Abd-Alsabour, "On the role of dimensionality reduction." *J. Comput.*, vol. 13, no. 5, pp. 571–579, 2018.

[135] N. Sharma, V. Jain, and A. Mishra, "An analysis of convolutional neural networks for image classification," *Procedia computer science*, vol. 132, pp. 377–384, 2018.

[136] N. Chumerin and M. M. Van Hulle, "Comparison of two feature extraction methods based on maximization of mutual information," in *2006 16th IEEE signal processing society workshop on machine learning for signal processing*. IEEE, 2006, pp. 343–348.

[137] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction," *Journal of Applied Science and Technology Trends*, vol. 1, no. 2, pp. 56–70, 2020.

[138] J. Wang and M. Wang, "Review of the emotional feature extraction and classification using eeg signals," *Cognitive Robotics*, 2021.

[139] F. Ahmed, H. Bari, and E. Hossain, "Person-independent facial expression recognition based on compound local binary pattern (clbp)." *Int. Arab J. Inf. Technol.*, vol. 11, no. 2, pp. 195–203, 2014.

[140] D. Gabor, "Theory of communication. part 1: The analysis of information," *Journal of the Institution of Electrical Engineers-part III: radio and communication engineering*, vol. 93, no. 26, pp. 429–441, 1946.

[141] N. Dalal, "Hostograms of oriented gradients for human detection," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2005*, vol. 1, 2005, pp. 886–893.

[142] H. Wei and Q. Peng, "A block-wise frame difference method for real-time video motion detection," *International Journal of Advanced Robotic Systems*, vol. 15, no. 4, pp. 1–13, 2018.

[143] P. N. Huu and T. Phung Ngoc, "Hand gesture recognition algorithm using svm and hog model for control of robotic system," *Journal of Robotics*, vol. 2021, 2021.

[144] O. S. Ekundayo and S. Viriri, "Facial expression recognition: A review of trends and techniques," *IEEE Access*, vol. 9, pp. 136 944–136 973, 2021.

[145] S. Eckstein, A. Iske, and M. Trabs, "Dimensionality reduction and wasserstein stability for kernel regression," *arXiv preprint arXiv:2203.09347*, 2022.

[146] K. Raju, B. Chinna Rao, K. Saikumar, and N. Lakshman Pratap, "An optimal hybrid solution to local and global facial recognition through machine learning," in *A Fusion of Artificial Intelligence and Internet of Things for Emerging Cyber Systems*. Springer, 2022, pp. 203–226.

[147] A. S. Alharthi, S. U. Yunas, and K. B. Ozanyan, "Deep learning for monitoring of human gait: A review," *IEEE Sensors Journal*, vol. 19, no. 21, pp. 9575–9591, 2019.

[148] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," *arXiv preprint arXiv:1705.01842*, 2017.

[149] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5455–5516, 2020.

[150] R. Adrian, "Deep learning for computer vision with python," 2017.

[151] S. Targ, D. Almeida, and K. Lyman, "Resnet in resnet: Generalizing residual architectures," *arXiv preprint arXiv:1603.08029*, 2016.

[152] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[153] J. Wei, "Vgg neural networks: The next step after alexnet," *Towardsdatascience. com*, 2019.

[154] A. Khan, A. Chefranov, and H. Demirel, "Building discriminative features of scene recognition using multi-stages of inception-resnet-v2," *Applied Intelligence*, pp. 1–19, 2023.

[155] X. Xia, C. Xu, and B. Nan, "Inception-v3 for flower classification," in *2017 2nd international conference on image, vision and computing (ICIVC)*. IEEE, 2017, pp. 783–787.

[156] S. Wang, M. Xu, Y. Sun, G. Jiang, Y. Weng, X. Liu, G. Zhao, H. Fan, J. Li, C. Zou *et al.*, "Improved single shot detection using densenet for tiny target detection," *Concurrency and Computation: Practice and Experience*, vol. 35, no. 2, p. e7491, 2023.

[157] Y. Nan, J. Ju, Q. Hua, H. Zhang, and B. Wang, "A-mobilenet: An approach of facial expression recognition," *Alexandria Engineering Journal*, vol. 61, no. 6, pp. 4435–4444, 2022.

[158] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[159] D. Theckedath and R. Sedamkar, "Detecting affect states using vgg16, resnet50 and se-resnet50 networks," *SN Computer Science*, vol. 1, no. 2, pp. 1–7, 2020.

[160] S. Rajpal, N. Lakhyani, A. K. Singh, R. Kohli, and N. Kumar, "Using handpicked features in conjunction with resnet-50 for improved detection of covid-19 from chest x-ray images," *Chaos, Solitons & Fractals*, vol. 145, p. 110749, 2021.

[161] Q. A. Al-Haija and A. Adebanjo, "Breast cancer diagnosis in histopathological images using resnet-50 convolutional neural network," in *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE, 2020, pp. 1–7.

[162] S. Gupta, P. Kumar, and R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," *Multimedia Tools and Applications*, pp. 1–30, 2022.

[163] P. Smith and C. Chen, "Transfer learning with deep cnns for gender recognition and age estimation," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 2564–2571.

[164] Y. Xie, W. Tian, and T. Ma, "A transfer learning approach to compound facial expression recognition," in *2020 4th International Conference on Advances in Image Processing*, 2020, pp. 95–101.

[165] X. Wang, C. Jin, W. Liu, M. Hu, L. Xu, and F. Ren, "Feature fusion of hog and wld for facial expression recognition," in *Proceedings of the 2013 IEEE/SICE International Symposium on System Integration*. IEEE, 2013, pp. 227–232.

[166] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[167] T. Song, W. Zheng, P. Song, and Z. Cui, "Eeg emotion recognition using dynamical graph convolutional neural networks," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 532–541, 2018.

[168] A. Graves, C. Mayer, M. Wimmer, J. Schmidhuber, and B. Radig, "Facial expression recognition with recurrent neural networks," in *Proceedings of the International Workshop on Cognition for Technical Systems*, 2008.

[169] F. Manessi, A. Rozza, and M. Manzo, "Dynamic graph convolutional networks," *Pattern Recognition*, vol. 97, p. 107000, 2020.

[170] J. Cheng, M. Chen, C. Li, Y. Liu, R. Song, A. Liu, and X. Chen, "Emotion recognition from multi-channel eeg via deep forest," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 453–464, 2020.

[171] D. Şen and M. Sert, "Continuous valence prediction using recurrent neural networks with facial expressions and eeg signals," in *2018 26th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2018, pp. 1–4.

[172] H. Mayer, F. Gomez, D. Wierstra, I. Nagy, A. Knoll, and J. Schmidhuber, "A system for robotic heart surgery that learns to tie knots using recurrent neural networks," *Advanced Robotics*, vol. 22, no. 13-14, pp. 1521–1537, 2008.

[173] J. Schmidhuber, D. Wierstra, and F. J. Gomez, "Evolino: Hybrid neuroevolution/optimal linear search for sequence prediction," in *Proceedings of the 19th International Joint Conferenceon Artificial Intelligence (IJCAI)*, 2005.

[174] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with lstm recurrent networks," *Journal of machine learning research*, vol. 3, no. Aug, pp. 115–143, 2002.

[175] D. Eck and J. Schmidhuber, "Learning the long-term structure of the blues," in *International Conference on Artificial Neural Networks*. Springer, 2002, pp. 284–289.

[176] J. Schmidhuber, F. Gers, and D. Eck, "Learning nonregular languages: A comparison of simple recurrent networks and lstm," *Neural computation*, vol. 14, no. 9, pp. 2039–2041, 2002.

[177] A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks," in *Advances in neural information processing systems*, 2009, pp. 545–552.

[178] B. Qolomany, A. Al-Fuqaha, A. Gupta, D. Benhaddou, S. Alwajidi, J. Qadir, and A. C. Fong, "Leveraging machine learning and big data for smart buildings: A comprehensive survey," *IEEE Access*, vol. 7, pp. 90 316–90 356, 2019.

[179] S. Hochreiter, M. Heusel, and K. Obermayer, "Fast model-based protein homology detection without alignment," *Bioinformatics*, vol. 23, no. 14, pp. 1728–1736, 2007.

[180] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long short term memory networks for anomaly detection in time series," in *Proceedings*. Presses universitaires de Louvain, 2015, p. 89.

[181] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "A comparative analysis of forecasting financial time series using arima, lstm, and bilstm," *arXiv preprint arXiv:1911.09512*, 2019.

[182] ——, "The performance of lstm and bilstm in forecasting time series," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 3285–3292.

[183] W. Ullah, A. Ullah, I. U. Haq, K. Muhammad, M. Sajjad, and S. W. Baik, "Cnn features with bi-directional lstm for real-time anomaly detection in surveillance networks," *Multimedia tools and applications*, vol. 80, pp. 16 979–16 995, 2021.

[184] V. M. Joshi and R. B. Ghongade, "Idea: Intellect database for emotion analysis using eeg signal," *Journal of King Saud University-Computer and Information Sciences*, 2020.

[185] C. Tang, P. Xu, Z. Luo, G. Zhao, and T. Zou, "Automatic facial expression analysis of students in teaching environments," in *Chinese Conference on Biometric Recognition*. Springer, 2015, pp. 439–447.

[186] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.

[187] J. K. Tarus, Z. Niu, and G. Mustafa, "Knowledge-based recommendation: a review of ontology-based recommender systems for e-learning," *Artificial intelligence review*, vol. 50, no. 1, pp. 21–48, 2018.

[188] D. O. Anne, A. Chepkemoi, E. Maina *et al.*, "Automatic student affective state detection from plain text," *International Journal of Formal Sciences: Current and Future Research Trends*, vol. 10, no. 01, pp. 41–60, 2021.

[189] M. Villaret *et al.*, "Affective state-based framework for e-learning systems," in *Artificial Intelligence Research and Development: Proceedings of the 23rd International Conference of the Catalan Association for Artificial Intelligence*, vol. 339. IOS Press, 2021, p. 357.

[190] L. Learning(n.d.), "What is feedback and why is it important?" https://leolearning.com/blog/learning-content/what-is-feedback-and-why-is-it-important/, accessed: 2022-08-08.

[191] "Feedback and reporting," https://www.education.vic.gov.au/school/teachers/teaching resources/practice/Pages/insight-feedback.aspx, accessed:2022-08-08.

[192] A. P. Cavalcanti, A. Barbosa, R. Carvalho, F. Freitas, Y.-S. Tsai, D. Gašević, and R. F. Mello, "Automatic feedback in online learning environments: A systematic literature review," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100027, 2021.

[193] J. T. Richardson, "Instruments for obtaining student feedback: A review of the literature," *Assessment & evaluation in higher education*, vol. 30, no. 4, pp. 387–415, 2005.

[194] A. Taskiran and M. Yazici, "Formative feedback in online distance language learning: Boosting motivation with automated feedback," in *Motivation, Volition, and Engagement in Online Distance Learning*. IGI Global, 2021, pp. 100–125.

[195] M. Sailer, E. Bauer, R. Hofmann, J. Kiesewetter, J. Glas, I. Gurevych, and F. Fischer, "Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning," *Learning and Instruction*, p. 101620, 2022.

[196] C. J. Villagrá-Arnedo, F. J. Gallego-Durán, F. Llorens-Largo, P. Compañ-Rosique, R. Satorre-Cuerda, and R. Molina-Carmona, "Improving the expressiveness of black-box models for predicting student performance," *Computers in Human Behavior*, vol. 72, pp. 621–631, 2017.

[197] M. Berland, R. S. Baker, and P. Blikstein, "Educational data mining and learning analytics: Applications to constructionist research," *Technology, Knowledge and Learning*, vol. 19, no. 1, pp. 205–220, 2014.

[198] R. G. Lord, J. M. Diefendorff, A. M. Schmidt, and R. J. Hall, "Self-regulation at work," *Annual review of psychology*, vol. 61, no. 1, pp. 543–568, 2010.

[199] G. Prince, *A dictionary of narratology*. U of Nebraska Press, 2003.

[200] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *Journal of Artificial Intelligence Research*, vol. 61, pp. 65–170, 2018.

[201] C. van der Lee, E. Krahmer, and S. Wubben, "Pass: A dutch data-to-text system for soccer, targeted towards specific audiences," in *Proceedings of the 10th International Conference on Natural Language Generation*, 2017, pp. 95–104.

[202] C. Derici, K. Celik, E. Kutbay, Y. Aydın, T. Güngör, A. Özgür, and G. Kartal, "Question analysis for a closed domain question answering system," in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2015, pp. 468–482.

[203] C. Asaju and H. Vadapalli, "A temporal approach to facial emotion expression recognition," in *Southern African Conference for Artificial Intelligence Research*. Springer, 2021, pp. 274–286.

[204] Khanna.C, "Question Answering with a fine-tuned BERT," https://towardsdatascience.com/question-answering-with-a-fine-tuned-bert-bc4dafd45626, accessed:2022-09-20.

[205] S. P. Lende and M. Raghuwanshi, "Question answering system on education acts using nlp techniques," in *2016 world conference on futuristic trends in research and innovation for social welfare (Startup Conclave)*. IEEE, 2016, pp. 1–6.

[206] S. Hussain, O. Ameri Sianaki, and N. Ababneh, "A survey on conversational agents/chatbots classification and design techniques," in *Workshops of the International Conference on Advanced Information Networking and Applications*. Springer, 2019, pp. 946–956.

[207] M. Binkis, R. Kubiliūnas, R. Sturienė, T. Dulinskienė, T. Blažauskas, and V. Jakštienė, "Rule-based chatbot integration into software engineering course," in *International Conference on Information and Software Technologies*. Springer, 2021, pp. 367–377.

[208] K. Ramesh, S. Ravishankaran, A. Joshi, and K. Chandrasekaran, "A survey of design techniques for conversational agents," in *Information, Communication and Computing Technology: Second International Conference, ICICCT 2017, New Delhi, India, May 13, 2017, Revised Selected Papers*. Springer, 2017, pp. 336–350.

[209] Y. Narekar and S. Chavhan, "Ai based chatbot for the analysis of admission eligibility," vol. IX Issue V, 2020.

[210] Y. W. Chandra and S. Suyanto, "Indonesian chatbot of university admission using a question answering system based on sequence-to-sequence model," *Procedia Computer Science*, vol. 157, pp. 367–374, 2019.

[211] K. Hwerbi, "An ontology-based chatbot for crises management: use case coronavirus," *arXiv preprint arXiv:2011.02340*, 2020.

[212] S. Ilić, R. Nakano, and I. Hajnal, "Designing dialogue systems: A mean, grumpy, sarcastic chatbot in the browser," *arXiv preprint arXiv:1909.09531*, 2019.

[213] E. Amer, A. Hazem, O. Farouk, A. Louca, Y. Mohamed, and M. Ashraf, "A proposed chatbot framework for covid-19," in *2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, 2021, pp. 263–268.

[214] M. M. Sathik and G. Sofia, "Identification of student comprehension using forehead wrinkles," in *2011 International Conference on Computer, Communication and Electrical Technology (ICCCET)*. IEEE, 2011, pp. 66–70.

[215] A. Pise, H. Vadapalli, and I. Sanders, "Facial emotion recognition using temporal relational network: an application to e-learning," *Multimedia Tools and Applications*, pp. 1–21, 2020.

[216] S. Hinduja and S. Canavan, "Real-time action unit intensity detection," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 916–916.

[217] Y. Wei, H. Wang, M. Sun, and J. Liu, "Attention based relation network for facial action units recognition," *arXiv preprint arXiv:2210.13988*, 2022.

[218] J. Liao, Y. Liang, and J. Pan, "Deep facial spatiotemporal network for engagement prediction in online learning," *Applied Intelligence*, pp. 1–13, 2021.

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

[220] H. L. Fwa, "Fine-grained detection of academic emotions with spatial temporal graph attention networks using facial landmarks," 2022.

[221] P. Marcelino, "Transfer learning from pre-trained models: How to solve any image classification problem quickly and easily."

[222] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation*, vol. 29, no. 9, pp. 2352–2449, 2017.

[223] M. Rahman, Y. Watanobe, K. Nakamura *et al.*, "A bidirectional lstm language model for code evaluation and repair," *Symmetry*, vol. 13, no. 2, p. 247, 2021.

[224] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *2013 IEEE workshop on automatic speech recognition and understanding*. IEEE, 2013, pp. 273–278.

[225] R. Horev, "Bert explained: State of the art language model for nlp," *Towards Data Science*, vol. 10, 2018.

[226] J. Kaiser, M. M. Crespo-Llado, C. Turati, and E. Geangu, "The development of spontaneous facial responses to others' emotions in infancy: An emg study," *Scientific Reports*, vol. 7, no. 1, pp. 1–10, 2017.

[227] R. Paranduk and Y. Karisi, "The effectiveness of non-verbal communication in teaching and learning english: a systematic review," *Journal of English Culture, Language, Literature and Education*, vol. 8, no. 2, pp. 140–154, 2020.

[228] P. E. Ekman and R. J. Davidson, *The nature of emotion: Fundamental questions.* Oxford University Press, 1994.

[229] J. Panksepp, "Toward a general psychobiological theory of emotions," *Behavioral and Brain sciences*, vol. 5, no. 3, pp. 407–422, 1982.

[230] A. R. Damasio, "Emotions and feelings," in *Feelings and emotions: The Amsterdam symposium.* Cambridge University Press Cambridge, UK, 2004, pp. 49–57.

[231] I. G. (n.d.), "What is affective state," https://www.igi-global.com/dictionary/affective-state/42305, accessed:2022-09-23.

[232]

[233] C. Odo, "Adapting learning activities selection in an intelligent tutoring system to affect," in *Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part II 19.* Springer, 2018, pp. 521–525.

[234] E. T. Welsh, C. R. Wanberg, K. G. Brown, and M. J. Simmering, "E-learning: emerging uses, empirical results and future directions," *international Journal of Training and Development*, vol. 7, no. 4, pp. 245–258, 2003.

[235] L. Epignosis, "elearning 101 concepts, trends, applications," 2014.

[236] N. A. Dahan, M. Al-Razgan, A. Al-Laith, M. A. Alsoufi, M. S. Al-Asaly, and T. Alfakih, "Metaverse framework: A case study on e-learning environment (elem)," *Electronics*, vol. 11, no. 10, p. 1616, 2022.

[237] J. Phillips, "Nonverbal communication: An essential skill in the workplace," *Australian Medical Record Journal*, vol. 23, no. 4, pp. 132–134, 1993.

[238] D. Phutela, "The importance of non-verbal communication," *IUP Journal of Soft Skills*, vol. 9, no. 4, p. 43, 2015.

[239] J. Brownlee, *Long Short-term Memory Networks with Python: Develop Sequence Prediction Models with Deep Learning.* Jason Brownlee, 2017.

[240] A. Vidyarthi and N. Mittal, "Texture based feature extraction method for classification of brain tumor mri," *Journal of Intelligent & Fuzzy Systems*, vol. 32, no. 4, pp. 2807–2818, 2017.

[241] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, 2016.

[242] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[243] G. Philipp, D. Song, and J. G. Carbonell, "Gradients explode-deep networks are shallow-resnet explained," 2018.

[244] J. Peng, S. Kang, Z. Ning, H. Deng, J. Shen, Y. Xu, J. Zhang, W. Zhao, X. Li, W. Gong *et al.*, "Residual convolutional neural network for predicting response of transarterial chemoembolization in hepatocellular carcinoma from ct imaging," *European radiology*, vol. 30, no. 1, pp. 413–424, 2020.

[245] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[246] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with Long Short Term Memory," 1999.

[247] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[248] R. S. Roy and A. Anand, "Question answering for the curated web: Tasks and methods in qa over knowledge bases and text collections," *Synthesis Lectures onSynthesis Lectures on Information Concepts, Retrieval, and Services*, vol. 13, no. 4, pp. 1–194, 2021.

[249] H. Brown, K. Lee, F. Mireshghallah, R. Shokri, and F. Tramèr, "What does it mean for a language model to preserve privacy?" *arXiv preprint arXiv:2202.05520*, 2022.

[250] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.

[251] W. Yang, Y. Xie, A. Lin, X. Li, L. Tan, K. Xiong, M. Li, and J. Lin, "End-to-end open-domain question answering with bertserini," *arXiv preprint arXiv:1902.01718*, 2019.

[252] R. Joshi, "Accuracy, precision, recall & f1 score: Interpretation of performance measures," *Retrieved April*, vol. 1, no. 2018, p. 2016, 2016.

[253] M. Algarni, F. Saeed, T. Al-Hadhrami, F. Ghabban, and M. Al-Sarem, "Deep learning-based approach for emotion recognition using electroencephalography (eeg) signals using bi-directional long short-term memory (bi-lstm)," *Sensors*, vol. 22, no. 8, p. 2976, 2022.

[254] W. H. Abdulsalam, R. S. Alhamdani, and M. N. Abdullah, "Facial emotion recognition from videos using deep convolutional neural networks," *International Journal of Machine Learning and Computing*, vol. 9, no. 1, pp. 14–19, 2019.

[255] J. Hernandez, D. McDuff, A. Fung, M. Czerwinski *et al.*, "Deepfn: Towards generalizable facial action unit recognition with deep face normalization," *arXiv preprint arXiv:2103.02484*, 2021.

[256] V. Amal, S. Suresh, and G. Deepa, "Real-time emotion recognition from facial expressions using convolutional neural network with fer2013 dataset," in *Ubiquitous Intelligent Systems*. Springer, 2022, pp. 541–551.

[257] A. Boughida, M. N. Kouahla, and Y. Lafifi, "A novel approach for facial expression recognition based on gabor filters and genetic algorithm," *Evolving Systems*, pp. 1–15, 2021.

[258] I. Spivaka, S. Krepycha, O. Fedorova, and S. Spivakb, "Approach to recognizing of visualized human emotions for marketing decision making systems," 2021.

[259] K. Gasper, C. L. Danube, and D. Hu, "Making room for neutral affect: Evidence indicating that neutral affect is independent of and co-occurs with eight affective states," *Motivation and Emotion*, vol. 45, no. 1, pp. 103–121, 2021.

[260] W. Oliveira, "Towards automatic flow experience identification in educational systems: A human-computer interaction approach," in *Extended abstracts of the annual symposium on computer-human interaction in play companion extended abstracts*, 2019, pp. 41–46.

[261] D. J. Nicol and D. Macfarlane-Dick, "Formative assessment and self-regulated learning: A model and seven principles of good feedback practice," *Studies in higher education*, vol. 31, no. 2, pp. 199–218, 2006.

[262] F. Wolter and M. Zakharyaschev, "Qualitative spatio-temporal representation and reasoning: a computational perspective," *Exploring Artificial Intelligence in the New Millenium*, pp. 175–216, 2002.

[263] P. Khosravi-Kamrani, X. Qiao, G. Zanardi, C. A. Wiesen, G. Slade, and S. A. Frazier-Bowers, "A machine learning approach to determine the prognosis of patients with class iii malocclusion," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 161, no. 1, pp. e1–e11, 2022.

[264] X. Zhang, M. A. Hassan, and J. G. Prendergast, "Using machine learning to detect the differential usage of novel gene isoforms," *BMC bioinformatics*, vol. 23, no. 1, pp. 1–16, 2022.

[265] A. Adate, R. Saxena, and B. Gladys Gnana Kiruba, "Analysing image compression using generative adversarial networks," in *Soft Computing for Problem Solving*. Springer, 2019, pp. 425–432.

[266] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Transactions on Affective Computing*, pp. 1–12, 2022.

[267] P. Tasca, A. Hayes, and S. Liu, "The evolution of the bitcoin economy: Extracting and analyzing the network of payment relationships," *The Journal of Risk Finance*, vol. 19, no. 2, pp. 94–126, 2018.

[268] D. Koletsi, M. Solmi, N. Pandis, P. S. Fleming, C. U. Correll, and J. P. Ioannidis, "Most recommended medical interventions reach p< 0.005 for their primary outcomes in meta-analyses," *International journal of epidemiology*, vol. 49, no. 3, pp. 885–893, 2020.

[269] R. RK, S. Shanthakumar, P. Vykunth, and K. Sairamnath, "Real-time attention span tracking in online education," *arXiv preprint arXiv:2111.14707*, 2021.

[270] R. Rahul, S. Shanthakumar, P. Vykunth, and K. Sairamnath, "Real-time attention span tracking in online education," in *2020 IEEE MIT Undergraduate Research Technology Conference (URTC)*. IEEE, 2020, pp. 1–4.

[271] O. Mohamad Nezami, M. Dras, L. Hamey, D. Richards, S. Wan, and C. Paris, "Automatic recognition of student engagement using deep learning and facial expression," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2019, pp. 273–289.

[272] P. Bhardwaj, P. Gupta, H. Panwar, M. K. Siddiqui, R. Morales-Menendez, and A. Bhaik, "Application of deep learning on student engagement in e-learning environments," *Computers & Electrical Engineering*, vol. 93, p. 107277, 2021.

[273] P. Sharma, S. Joshi, S. Gautam, S. Maharjan, V. Filipe, and M. J. Reis, "Student engagement detection using emotion analysis, eye tracking and head movement with machine learning," *arXiv preprint arXiv:1909.12913*, 2019.

appendix