KING KHALID UNIVERSITY

COLLEGE OF COMPUTER SCIENCE

EduSense

An Intelligent Classroom System for Detecting Student Confusion Using Computer Vision

By

Saeed Mohammed S Asiri **444810913**
Fahad Abdullah Ali AL-Qahtani **444802593**
Khalid Mushabbab Al-Dahwan **444803647**
Ahmad Turki Al Sultan **444803284**
Basil Hasan Al Muawwadh **442811409**

Supervised by:
Anand Deva Durai C

ACKNOWLEDGMENT

First and foremost, all praise and thanks be to Allah, whose guidance and blessings have enabled us to complete this graduation project successfully.

We would like to express our deepest gratitude to our supervisor, **Dr. Anand Deva Durai C** , for his continuous support, valuable guidance, and encouragement throughout this project. His insightful feedback and expertise have been instrumental in shaping the outcome of this work.

Our sincere appreciation extends to the College of Computer Science at **King Khalid University** for providing the academic environment and resources that made this project possible.

We would also like to thank our colleagues and friends for their constant help, motivation, and collaboration during every stage of development.

Finally, our heartfelt thanks go to our beloved families, whose patience, prayers, and unconditional support have been our greatest source of strength and inspiration throughout our academic journey.

ABSTRACT

The growing shift toward online education has created new challenges for instructors, who often lack real-time insight into student comprehension during video-based learning. Without the immediate feedback present in physical classrooms, it becomes difficult to identify when learners struggle with complex topics. This project introduces **EduSense**, an intelligent system that leverages computer vision and machine learning to detect moments of confusion in online settings and automatically generate adaptive learning resources.

Using webcam input, EduSense captures learners' facial expressions, gaze direction, and head movements as they engage with educational videos. These multimodal cues are analyzed to estimate confusion levels over time, forming a **"confusion curve"** synchronized with the video timeline. Peaks in confusion are mapped to corresponding transcript segments and video frames, which are then transformed into **structured notes**, **interactive quizzes**, and—when programming content is detected—**Jupyter notebooks** containing runnable code.

To enhance interpretability and modeling precision, EduSense employs **Kolmogorov–Arnold Networks (KANs)** as an alternative to traditional deep learning models. KANs offer strong function approximation and improved transparency, enabling better understanding of the nonlinear relationships between micro-expressions, gaze dynamics, and cognitive states.

Overall, EduSense provides educators with actionable insights into student engagement and confusion during online learning, establishing a foundation for future research in **AI-driven learning analytics** and **interpretable educational technologies**.

.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

The rapid growth of online learning has expanded educational access but eliminated real-time feedback that helps instructors gauge student understanding. In traditional classrooms, teachers rely on expressions and gaze to identify confusion, while online video-based environments create one-way communication that obscures learning difficulties. This gap leads to reduced engagement and high dropout rates in digital education.

The **general problem** addressed in this project is the absence of intelligent, interpretable systems that detect and respond to student confusion during online learning. Despite progress in affective computing, current methods remain limited by poor interpretability, low temporal precision, and lack of adaptive feedback capable of improving learner outcomes in real time.

This study explores three **hypotheses**:
(H1) Kolmogorov–Arnold Networks (KANs) provide accurate and interpretable confusion detection;
(H2) aligning confusion signals with video transcripts localizes difficult content; and
(H3) generating adaptive materials from these segments improves learning outcomes.

To test them, three **objectives** are set: implement and benchmark KANs (O1); develop a temporal analysis pipeline to map confusion to content (O2); and integrate large language models to generate notes and quizzes from confused moments (O3).

Completing these objectives will validate each hypothesis respectively—demonstrating interpretability gains (H1), diagnostic precision (H2), and educational value (H3). The **anticipated impact** is an integrated, privacy-conscious framework that detects, explains, and addresses confusion, transforming passive online learning into an adaptive, data-driven experience that enhances comprehension and retention. Ultimately, this work aims to narrow the feedback gap between learners and instructors, paving the way for more personalized and effective online education worldwide.

CHAPTER 2 REVIEW OF LITERATURE

2.1 INTRODUCTION

The proliferation of online and hybrid learning modalities has fundamentally transformed the educational landscape, creating unprecedented opportunities for scalable, accessible instruction. However, this shift has simultaneously introduced significant pedagogical challenges, particularly regarding real-time assessment of learner comprehension and engagement. Unlike traditional classroom settings where instructors can observe facial expressions, body language, and verbal cues to gauge understanding, digital learning environments often lack mechanisms for immediate feedback, making it difficult to identify when students encounter cognitive obstacles or confusion.

This literature review examines the evolution of affective computing and intelligent educational systems, with particular emphasis on confusion detection, multimodal affect recognition, and adaptive learning resource generation. The review is organized into several thematic sections: the historical development of emotion recognition systems (Section 2.2), advances in temporal modeling and attention mechanisms (Section 2.3), the unique characteristics of confusion as an epistemic emotion (Section 2.4), emerging neural architectures including Kolmogorov–Arnold Networks (Section 2.5), automated content generation for personalized learning (Section 2.6), and finally, a synthesis identifying critical research gaps that motivate the EduSense system (Section 2.7).

2.2 EVOLUTION OF EMOTION RECOGNITION SYSTEMS

2.2.1 Early Approaches: Handcrafted Features and Classical Methods

The field of automated emotion recognition emerged in the 1990s with the pioneering work of Ekman and Friesen, who developed the Facial Action Coding System (FACS)—a comprehensive taxonomy of facial muscle movements corresponding to emotional expressions. Early computational systems leveraged FACS units combined with classical machine learning techniques such as Support Vector Machines (SVMs) and Hidden Markov Models (HMMs) to classify basic emotions (happiness, sadness, anger, fear, surprise, and disgust). While these approaches demonstrated proof-of-concept viability, they were constrained by several limitations: reliance on controlled laboratory conditions, sensitivity to lighting variations and head pose, and inability to capture subtle or compound affective states.

2.2.2 Deep Learning Revolution: CNNs and Feature Learning

The advent of Convolutional Neural Networks (CNNs) marked a paradigm shift in emotion recognition. Models such as VGGNet, ResNet, and Inception architectures enabled automatic feature extraction from raw pixel data, eliminating the need for manual feature engineering. Benchmarking on datasets like FER2013, CK+, and AffectNet demonstrated substantial accuracy improvements over handcrafted approaches. However, CNN-based systems remained fundamentally limited by their frame-by-frame processing paradigm, treating each image independently without considering temporal dynamics—a critical oversight given that emotions unfold over time through micro-expressions and transitional states.

2.2.3 Commercial and Research Systems

Contemporary commercial platforms such as Affectiva and iMotions have operationalized emotion recognition for market research, user experience testing, and automotive safety applications. Affectiva's Emotion AI SDK utilizes deep learning models trained on millions of faces across diverse demographics to detect seven core emotions plus 20 facial expressions in real time. iMotions integrates facial expression analysis (FEA) with complementary biometric modalities including eye tracking, galvanic skin response, and electroencephalography (EEG), enabling multimodal affective measurement. While these systems achieve robust performance in controlled settings, their computational overhead, licensing costs, and generalizability to educational contexts remain practical barriers for widespread classroom deployment.

2.3 TEMPORAL MODELING AND ATTENTION MECHANISMS

2.3.1 Recurrent Architectures for Sequential Data

Recognizing the inherently temporal nature of emotional expressions, researchers have increasingly adopted Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and Gated Recurrent Units (GRUs) to model temporal dependencies in video sequences. These architectures maintain hidden states that encode historical information, enabling them to capture how expressions evolve and transition over time. Studies have demonstrated that LSTM-based models outperform frame-level CNNs on video-based emotion recognition tasks, particularly for detecting subtle affective shifts and transient micro-expressions.

2.3.2 Transformer-Based Temporal Attention

The introduction of Transformer architectures—originally developed for natural language processing—has revolutionized temporal sequence modeling across multiple domains,

including computer vision. Unlike RNNs, Transformers employ self-attention mechanisms that enable parallel processing of entire sequences while dynamically weighting the importance of different temporal positions. Vision Transformers (ViTs) and their variants have achieved state-of-the-art performance on video-based emotion recognition benchmarks by learning long-range temporal dependencies without the vanishing gradient problems inherent to recurrent architectures. Recent work has extended these models to multi-head attention configurations that jointly attend to spatial facial regions and temporal frames, yielding more nuanced representations of affective dynamics.

2.3.3 Application to Educational Contexts

The EmoAI Smart Classroom system exemplifies the application of advanced temporal modeling to educational affect recognition. Utilizing YOLOv8 for face detection combined with temporal analysis, the system monitors student engagement (boredom, confusion, frustration, and focus) in large offline classrooms. Despite achieving 74.5% precision and 65.3% mean Average Precision (mAP), the study identified critical limitations including computational latency, limited training data (707 annotated frames from 2.7 million extracted), and challenges with occlusions (e.g., students wearing glasses). These findings underscore the ongoing need for lightweight, real-time systems optimized for educational deployment.

2.4 CONFUSION DETECTION: AN EPISTEMIC EMOTION

2.4.1 Theoretical Foundations

Confusion occupies a unique position in the taxonomy of emotions. Unlike basic affective states such as happiness, sadness, or anger, confusion is classified as an epistemic emotion—one that emerges during knowledge construction and problem-solving when learners encounter cognitive impasses or conflicting information. Theoretical frameworks in educational psychology, such as D'Mello and Graesser's cognitive-affective model of learning, suggest that moderate confusion can be productive when it stimulates reflection and inquiry. However, unresolved or prolonged confusion can lead to frustration and disengagement, making its detection critical for timely instructional support in digital learning environments.

2.4.2 Behavioral Manifestations

Confusion can be inferred from a range of **visual behavioral cues** observable through a standard camera. Facial expressions often include furrowed brows, narrowed or squinting

eyes, asymmetric lip movements, and micro-expressions indicating uncertainty. Gaze behavior shows longer fixation durations, increased regression to previously viewed areas, and reduced blink rates, reflecting heightened cognitive load. Head movements may become irregular, including frequent tilting or subtle shaking. These cues are typically subtle and transient, requiring computer vision systems with adequate temporal sensitivity to detect them reliably.

2.4.3 Prior Work in Confusion Detection

Several studies have explored the automatic detection of confusion using visual data. Peng and Nagao (2021) developed a classroom analysis framework that classified states such as boredom, concentration, and confusion based on facial and gaze patterns captured from video streams. Their system achieved high accuracy but relied on multiple hardware devices that limited scalability. Zheng et al. (2020) used deep learning to infer engagement levels from physical behaviors like hand-raising or head position in online classes. While effective for gross engagement estimation, their model did not explicitly target confusion or capture fine-grained facial dynamics.

The **DAiSEE (Dataset for Affective States in E-Environments)** remains one of the most influential datasets in this domain, containing over 9,000 annotated video clips labeled for boredom, confusion, engagement, and frustration. Studies leveraging DAiSEE have employed CNNs, LSTMs, and hybrid temporal models, achieving accuracies around 77%. Despite these advances, challenges persist—particularly in improving interpretability, achieving real-time inference, and generalizing across diverse learners and lighting conditions.

2.5 KOLMOGOROV–ARNOLD NETWORKS: AN EMERGING PARADIGM

2.5.1 Theoretical Background

Kolmogorov–Arnold Networks (KANs) represent a novel neural architecture grounded in the Kolmogorov–Arnold representation theorem, which states that any continuous multivariate function can be decomposed into a finite composition of continuous univariate functions and addition operations. Unlike traditional neural networks that employ fixed activation functions (ReLU, sigmoid, tanh) applied uniformly across neurons, KANs parameterize activation functions themselves as learnable univariate splines positioned on network edges rather than nodes. This architectural innovation

enables KANs to approximate complex functions with greater parameter efficiency and interpretability.

2.5.2 Advantages Over Conventional Deep Learning

Empirical studies have demonstrated that KANs achieve comparable or superior accuracy to Multi-Layer Perceptrons (MLPs) and CNNs on various regression and classification tasks while utilizing significantly fewer parameters—often by an order of magnitude. This compactness translates to reduced computational requirements, faster inference, and lower memory footprints—critical advantages for real-time educational applications. Additionally, KANs exhibit improved calibration, meaning their confidence estimates more accurately reflect true prediction probabilities. This property is particularly valuable in educational contexts where uncertain predictions should trigger alternative instructional strategies rather than erroneous interventions.

2.5.3 Interpretability and Educational Applications

Perhaps most significantly, KANs offer enhanced interpretability compared to black-box deep learning models. The learnable univariate activation functions can be visualized and analyzed to reveal which input features contribute most strongly to predictions and how they interact through compositional structure. In the context of confusion detection, this transparency could illuminate which specific combinations of facial action units, gaze metrics, and temporal patterns most reliably indicate cognitive struggle—insights that could inform both model refinement and pedagogical theory. Despite these promising attributes, no existing literature has applied KANs to affective computing or confusion detection, representing a significant research opportunity.

2.6 AUTOMATED LEARNING RESOURCE GENERATION

2.6.1 Intelligent Tutoring Systems and Adaptive Learning

Intelligent Tutoring Systems (ITSs) have evolved from rule-based expert systems to data-driven adaptive platforms that personalize content, pacing, and pedagogical strategies based on learner models. Early systems like AutoTutor employed natural language processing to engage students in dialogue-based learning, detecting affective states through linguistic cues and adapting feedback accordingly. Modern ITSs increasingly incorporate multimodal sensing and machine learning to infer knowledge states, predict performance, and recommend learning paths.

2.6.2 Micro-Learning and Just-in-Time Support

The micro-learning paradigm emphasizes delivering concise, focused learning resources (short videos, flashcards, practice problems) precisely when learners encounter difficulty. Research in cognitive load theory and multimedia learning suggests that targeted, timely interventions are more effective than overwhelming learners with comprehensive but poorly-timed content. Systems that automatically generate micro-resources from longer instructional videos—extracting key segments, summarizing concepts, and creating practice exercises—represent an emerging frontier in adaptive learning technologies.

2.6.3 Large Language Models for Content Generation

The recent advent of Large Language Models (LLMs) such as GPT-4, Claude, and Gemini has dramatically expanded possibilities for automated educational content creation. These models can generate structured notes, explanations, quiz questions, and even executable code from textual descriptions or video transcripts. When integrated with confusion detection systems, LLMs enable closed-loop adaptive learning: detect confusion → identify problematic content → generate personalized clarification resources → deliver intervention → reassess understanding. Despite their potential, challenges remain regarding factual accuracy, pedagogical appropriateness, and alignment with learning objectives—necessitating human oversight and validation.

2.7 RESEARCH GAPS

1. **Lack of Lightweight Real-Time Confusion Detection Systems**
   Many existing emotion recognition systems are designed mainly for accuracy and rely on heavy computational resources, which makes them unsuitable for real-time classroom environments. There is a need for a lightweight confusion detection system that can operate efficiently on regular laptops or webcams without requiring high-end hardware.

2. **Limited Interpretability of Deep Learning Models**
   Most current deep learning models act as “black boxes,” providing little explanation of how they reach their predictions. In educational settings, instructors and learners need transparent and explainable systems to build trust and understand how confusion is detected.

3. **Insufficient Temporal Precision**
   Many existing systems detect emotions for entire sessions or videos instead of identifying the exact moments when confusion occurs. For teachers, knowing that a student was confused during a 45-second segment of a lecture is far more useful than a general label for the whole video.

4. **Lack of Automatic Feedback or Intervention**
   Although some systems can detect confusion, very few can respond automatically to it. There is a gap in developing systems that can transform detected confusion into helpful learning resources or feedback for the student.

5. **Unexplored Use of Kolmogorov–Arnold Networks (KANs)**
   Kolmogorov–Arnold Networks (KANs) are a new type of neural network known for being efficient and interpretable. However, they have not yet been applied to confusion detection, and their potential advantages in this area remain unexplored.

2.8 SUMMARY AND CONNECTION TO THE PROPOSED SYSTEM

The reviewed literature provides a strong foundation for developing intelligent educational systems that can detect and respond to student confusion in real time. Early emotion recognition systems proved the feasibility of using visual signals for affective analysis but often lacked the accuracy and temporal precision needed for realistic classroom environments. The introduction of deep learning, particularly temporal models such as LSTMs and Transformers, has improved performance on video-based emotion recognition tasks. Educational datasets like **DAiSEE** have supported this progress by offering labeled data for classroom scenarios, though challenges remain regarding computational efficiency, model interpretability, and data privacy.

Recent studies highlight the potential of **Kolmogorov–Arnold Networks (KANs)** as efficient and interpretable neural architectures. Their ability to balance accuracy and explainability makes them suitable for educational applications that require both transparency and low-latency performance. Additionally, recent advances in **Large Language Models (LLMs)** open opportunities for generating personalized learning materials automatically, bridging the gap between confusion detection and instructional support.

The **EduSense system** proposed in this project directly addresses the gaps identified in the literature by focusing on the following aspects:

- **Camera-Based Confusion Detection:**
  Analyzing facial expressions, gaze direction, and head movements captured through a single webcam to identify confusion with fine temporal accuracy.

- **Kolmogorov–Arnold Networks (KANs):**
  Implementing KANs as an interpretable and computationally efficient alternative

to conventional deep learning models, allowing transparent decision-making and smooth deployment on standard devices.

- **Temporal Synchronization:**
  Generating a confusion curve aligned with the video timeline to pinpoint the exact segments where learners struggle the most.

- **Automated Resource Generation:**
  Using LLMs to convert the detected confused segments into structured notes and short quizzes, closing the loop between detection and adaptive learning.

By combining these components, **EduSense** provides a practical and interpretable framework for confusion-aware online learning. The system not only addresses immediate educational challenges but also contributes valuable insights and tools for future research in learning analytics and explainable artificial intelligence.