Chapter 3

Overview of the Proposed Research Methodology for Addressing the Problem Statement

3.1 Introduction

The main objective of this thesis is to design a computationally efficient deep learning model for FER to analyze learner engagement in mulsemedia-synchronized learning environment. Conventional FER models face significant challenges, such as difficulties in extracting features from spatiotemporal data when deploying models in real-time environments. This is due to dynamic facial movements, whereas most existing FER models are trained on static facial expression images. Another important issue is the limited sample count for certain expressions in existing FER datasets, leading to class imbalance and overfitting when training a model with insufficient FER samples in a supervised learning approach. Additionally, conventional CNN-based FER models tend to increase layer depth to improve accuracy, which can lead to increased computational costs. These challenges result in less effective performance in real-time FER. These problems are thoroughly outlined in Chapter 2 through exhaustive systematic reviews. This chapter provides an overview of the proposed research methodology to address these challenges step-by-step to attain the main objectives.

Secondary Objectives:

- To experiment with various deep learning approaches (3D-CNN, LSTM, and various types of Autoencoder) for building effective way of FER systems in mulsemedia learning environments using a universal facial expression dataset.
- To investigate the effectiveness of learners' satisfaction with mulsemedia-synchronised content, fostering improved engagement and knowledge retention among learners.
- Maximize classification accuracy and detection rate in FER.
- Minimize prediction time for FER.
- Maximize precision and recall scores in emotion classification.

3.2 Steps Involved in Proposed Methodology

The proposed FER system for the learning environment consists of four primary steps, as depicted in Figure 3.1. To accomplish the main objective, the approach outlined in Figure 3.2 has been implemented.

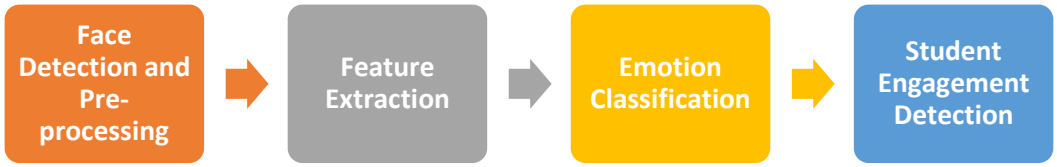

Figure 3.1. Process of Detecting Learner Engagement in Learning Environments

1. **Face Detection and Pre-processing**:

   This initial step involves detecting faces in images/video frames and pre-processing them by resizing, normalizing, and converting them to grayscale to ensure uniformity and enhance feature extraction efficiency.

2. **Feature Extraction:**

   Utilizing a deep learning approach, this step automates the feature extraction process, effectively overcoming the limitations associated with conventional handcrafted feature extraction methods. Deep learning models are adept at capturing intricate patterns and representations from facial images, which are crucial for accurate emotion recognition.

3. **Emotion Classification:**

   The final step involves classifying the extracted features into corresponding emotional labels. The system has been trained to recognize and categorize real-time facial expressions such as neutral, happy, surprise, and sleep. These expressions were collected from students in a controlled environment to ensure consistency and reliability in the training data. Furthermore, various metrics have been employed to assess the model's performance comprehensively:

   - **Accuracy**: Measures the proportion of correctly classified instances among the total instances.

- **Precision**: Indicates the proportion of true positive results among all positive results predicted by the model.
- **Recall**: Reflects the proportion of true positive results among all actual positive instances.
- **F1-Score**: The harmonic mean of precision and recall, providing a single metric that balances both concerns.
- **Confusion Matrices**: Present a detailed breakdown of true positives, false positives, true negatives, and false negatives for each class, offering insight into specific classification errors.
- **Receiver Operating Characteristic (ROC) Curve**: Illustrates the trade-off between the true positive rate and false positive rate, with the Area Under the Curve (AUC) indicating the model's ability to discriminate between classes.
- Additionally, to validate the generalization capability of the proposed model, further elucidated using XAI techniques. These techniques provide insights into the model's decision-making process, enhancing interpretability and helping to understand how the model arrives at its predictions.

4. **Student Engagement Detection:**

In digital learning environments, detecting learner engagement plays a crucial role in determining the most effective teaching styles or learning content to enhance students' positive emotions. Engagement refers to a student's level of involvement, interest, and motivation while interacting with educational material. It can be assessed through various methods, with FER emerging as a promising approach for measuring emotional engagement.

Some FER studies propose methods to map universal facial expressions to categories of learner engagement, such as engaged, neutral, and disengaged. These methods are based on affective mapping frameworks established in prior research, enabling a structured understanding of how facial expressions correlate with engagement levels. A notable study by Pise et al. [147] explored how learners' facial expressions could be used to interpret their emotional states during learning sessions with learning materials. They found this approach effective and also examined a standardized mechanism for mapping facial expressions to corresponding learning outcomes. Their findings confirmed that the system could accurately

detect facial expressions and associate each recognized emotion with its respective learning affect, consistent with the predefined mapping system.

Building on these methodologies, we evaluated the FER system within a learning environment, particularly in the context of integrating mulsemedia into learning content to enhance immersive learning experiences. This assessment aims to determine how effectively the FER system can gauge student engagement and emotional states while interacting with mulsemedia elements in educational content.

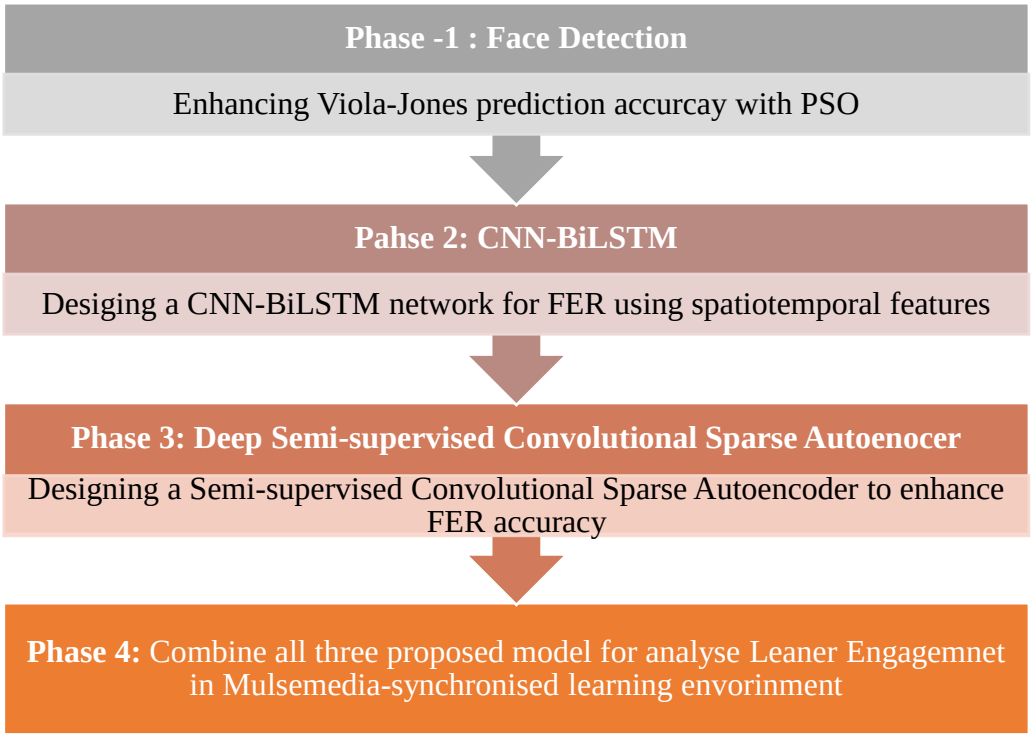

Figure 3.2. Phases of Research Work

3.2.1 Enhancing Viola-Jones Face Detection Algorithm Accuracy by PSO

In Contribution 1, we have improved the prediction accuracy of the Viola-Jones algorithm by incorporating the PSO algorithm. Face detection has become a key area of research in CV and DL and is also essential in FER during the pre-processing steps. Detecting faces in images and video sequences remains challenging due to factors like pose variation, changing illumination, occlusion, and scale differences. Although many face detection algorithms have been developed in deep learning, the Viola-Jones algorithm remains popular for real-time camera applications due to its simple yet effective approach. Traditionally, the

Viola-Jones algorithm uses AdaBoost to classify faces in images and videos. However, handling cluttered real-time facial images poses a challenge. AdaBoost must search through all possible thresholds for all samples to minimize training error when receiving features from Haar-like detectors. This exhaustive search process is time-consuming as it aims to find the best threshold values and optimize feature selection to build an efficient classifier for face detection. For this research gap, we incorporated PSO to improve its predictive accuracy, particularly in complex face images. We leverage PSO in two critical areas within the Viola-Jones framework.

- First, PSO is utilized to dynamically select optimal threshold values for feature selection, thereby enhancing computational efficiency.
- Second, we adapt the feature selection process using AdaBoost within the Viola-Jones algorithm, integrating PSO to identify the most discriminative features for constructing a robust classifier.

This approach significantly reduces the time required for feature selection and minimizes search complexity compared to conventional algorithms, particularly in challenging environments. The proposed method was evaluated on a comprehensive face detection benchmark dataset, achieving superior results, including an average true positive rate of 98.73% and a 2.1% higher average prediction accuracy compared to the conventional Viola-Jones algorithm and other contemporary state-of-the-art methods.

3.2.2 FER Using CNN-BiLSTM Architecture

In Contribution 2, facial expression recognition based on deep learning networks, the CNN has proven to be an effective method for extracting local spatial features of images. However, CNNs are typical feedforward deep networks with a monotonous, unidirectional structure, which limits their ability to recognize contextual timing information and restricts the algorithm's accuracy. On the other hand, LSTM and RNN networks are designed to capture temporal and global features. LSTM, an extended version of RNN, overcomes the vanishing gradient problem and enhances the capability to remember long sequence information.

In conventional hybrid CNN-LSTM or CNN-RNN models, video frames are processed one by one, predicting the emotion for each frame sequentially (potentially

skipping 4-5 frames to increase the fps). In this setup, the CNN extracts spatial features from individual frames, capturing local patterns like edges, textures, and shapes, while the LSTM processes the sequence of spatial features extracted by the CNN. However, this approach might not fully capture bidirectional temporal dependencies as it processes sequences in one direction.

To address this issue, we introduced a TimeDistributed layer, which allows the CNN to process each frame within the sequence, ensuring spatial features are consistently extracted across all frames. Additionally, we incorporated a Bi-LSTM network, which processes the sequence of features in both directions, capturing more comprehensive temporal information and leading to a more nuanced understanding of temporal patterns.

In this approach, a hyperparameter-tweaked VGG-19 skeleton is employed to automatically extract spatial features from a sequence of images, avoiding the shortcomings of conventional feature extraction methods. These features are then fed into a bidirectional Bi-LSTM to extract spatiotemporal features of time series in both directions, allowing for the recognition of emotion from a sequence of expressions.

The performance of the proposed method was evaluated using the CK+ benchmark as well as an In-house dataset captured from a designed Internet of Things (IoT) kit. The approach was verified through hold-out cross-validation techniques, achieving an accuracy of 92% on CK+ and 84% on the In-house dataset. The experimental results reveal that the proposed method outperforms baseline methods and state-of-the-art approaches. Furthermore, precision, recall, F1-score, and ROC curve metrics were used to evaluate the performance of the proposed system. In addition, an XAI approach, specifically the Grad-CAM technique, has been applied to validate the model's performance.

Moreover, a limitation of supervised learning is its reliance on labeled data, where each training example is paired with a correct label. In the context of FER, this involves datasets where each facial expression is tagged with its corresponding emotion. However, real-world FER datasets often suffer from imbalanced expression distributions, leading to overfitting, particularly when the dataset lacks diversity. While data augmentation can help mitigate overfitting by artificially expanding the training dataset, excessive augmentation risks data leakage. This occurs when the model performs well on the augmented data but

poorly on real-world data, where conditions differ from those artificially created. Therefore, extracting more discriminative features from limited samples remains a critical yet challenging task.

3.2.3 FER Implementation with Deep Semi-supervised Convolutional Sparse Autoencoder

In Contribution 3, an autoencoder is introduced as an Artificial Neural Network (ANN) primarily designed for unsupervised learning tasks. Its main objective is to create efficient data representations, commonly used for dimensionality reduction or feature extraction. The autoencoder consists of three main components:

- **Encoder**: This part of the network compresses the input data into a lower-dimensional representation, commonly known as the "bottleneck" or "latent space." It reduces the dimensionality of the input data while retaining crucial features.
- **Decoder**: This part reconstructs the original data from the compressed representation, aiming to approximate the original input as closely as possible.
- **Loss Function**: The autoencoder is trained to minimize the difference between the input data and its reconstruction, typically using a loss function like mean squared error.

In essence, autoencoders compress data into a compact form and then decode it, capturing the essential features in the process. This ability to represent data efficiently is useful for tasks such as data denoising, anomaly detection, and generating new data samples. In the context of FER, performance accuracy heavily relies on large, high-quality labeled facial expression datasets. A significant challenge with existing FER datasets is the variability in the number of samples and expressions for each emotion. Additionally, implementing FER using CNNs often necessitates many layers, which can extend training times and complicate the optimization of parameters. This complexity may hinder the creation of distinct facial expression patterns, leading to challenges in real-time emotion classification.

To address these issues, the proposed method, known as the Deep Semi-Supervised Convolutional Sparse Autoencoder (DSCSA), seeks to enhance FER performance by requiring fewer labeled samples and parameters compared to the conventional approach. The

process begins with face detection using an improved Viola-Jones algorithm optimized with PSO, followed by normalization and resizing of pixel ranges for training. The core of the method is a Deep Convolutional Sparse Autoencoder, which is trained on unlabeled FER samples. By imposing sparsity on a hidden layer in the encoder, the model reduces overfitting and eliminates irrelevant features, leading to an accurate reconstruction of the input. To enable semi-supervised learning, a SoftMax function is added on top of the encoder to classify facial expressions using both labeled and unlabeled data. This approach was evaluated on in-house data as well as benchmark datasets CK+ and JAFFE, achieving notable accuracies of 98.06%, 98.98%, and 93.10%, respectively. The results were analyzed using established state-of-the-art techniques. Additionally, XAI methods, including Grad-CAM and image-LIME, were employed to interpret the model's performance and prediction outcomes.

3.2.4 Analyse Learner's Engagement using FER in Mulsemedia-synchronized Learning Environment

In Contribution 4, we have designed a Mulsemedia-synchronized web portal for delivering learning content enhanced with mulsemedia effects. Conventional digital content primarily engages only two human senses: sight and sound. By incorporating additional sensory stimuli—such as olfactory (smell), haptic (touch), and gustatory (taste) feedback—into digital content, we can significantly enhance the user learning experience.

To analyze learners' engagement using FER in a mulsemedia-synchronized learning environment, we designed an IoT-based platform leveraging affordable components like cooling fans, humidifiers, and haptic devices to create a multisensory learning experience. This platform integrates sensory effects, such as the aroma of rosemary for rosemary-based audiovisual content and vibrotactile and airflow effects synchronized with thunder and lightning learning materials. For the study, 70 science degree students participated, divided equally into experimental and control groups (CG). The experimental group (EG) experienced the mulsemedia-enhanced content, while the control group engaged with conventional learning methods. Learners' engagement and the quality of their learning experiences were assessed through a self-reported questionnaire. The results demonstrated that mulsemedia-based learning significantly enhanced learning outcomes compared to traditional methods. Additionally, it increased enjoyment levels and provided a heightened

sense of reality in the learning environment, showcasing the potential of integrating mulsemedia to improve learner engagement and satisfaction.

Secondly, this approach, combined with the FER system as shown in Figure 3.4, assesses the engagement levels of 20 participants in the mulsemedia learning environment. The system begins by capturing an input video sequence of the learner through a camera. The video is preprocessed by converting it to grayscale to reduce complexity. Face detection is then performed using the Viola-Jones algorithm, optimized with PSO for improved accuracy. If no face is detected, the system continues searching in subsequent frames. Detected faces are resized to a standardized 48 x 48 resolution and normalized. From the processed frames, ten samples are selected for spatiotemporal features analysis, while a single sample is extracted for spatial analysis. These samples undergo feature extraction using two approaches: CNN-BiLSTM for spatiotemporal features and DSCSA for spatial features, leveraging pre-trained weights for accuracy. The extracted features are concatenated to classify facial expressions into categories such as Happy, Surprise, Fear, Neutral, Disgust, Sleepy, Sad, and Anger. These expressions are mapped to engagement levels using a proposed adaptive mapping unit: "Highly Engaged" (Happy, Surprise, Fear), "Engaged" (Neutral), and "Disengaged" (Disgust, Sad, Sleep, Anger). The system compares learner engagement in two groups: those exposed to mulsemedia (EG) and those without (CG). This framework aims to analyze how mulsemedia impacts engagement levels, enhancing the learning experience.The results indicate that mulsemedia in a learning environment significantly enhances learning outcomes compared to conventional approaches.

3.3 Consolidated View of the Proposed Methodology and Comparison with Other Methods

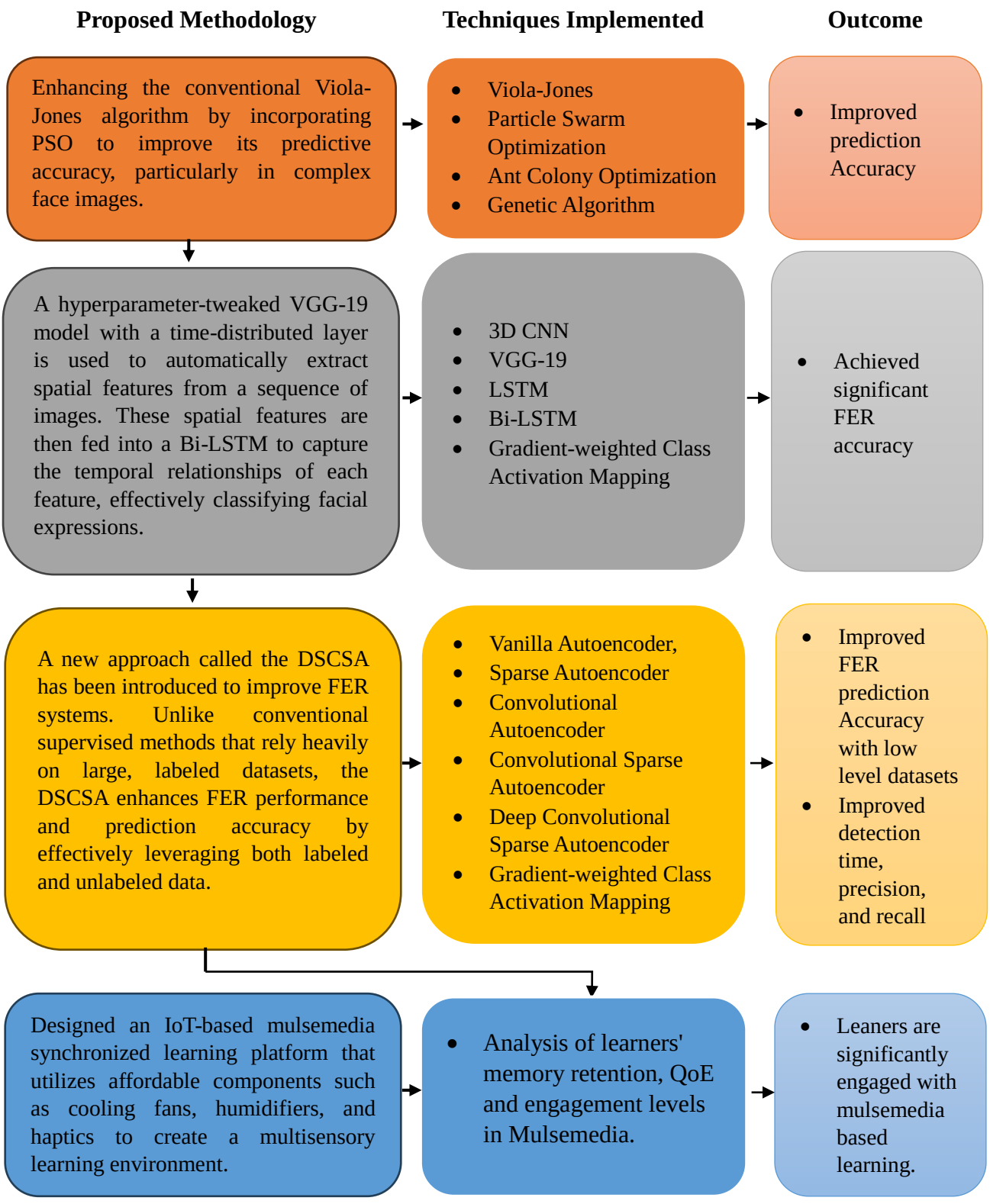

Figure 3.3 Overview of Proposed Methodology and Significant Contribution

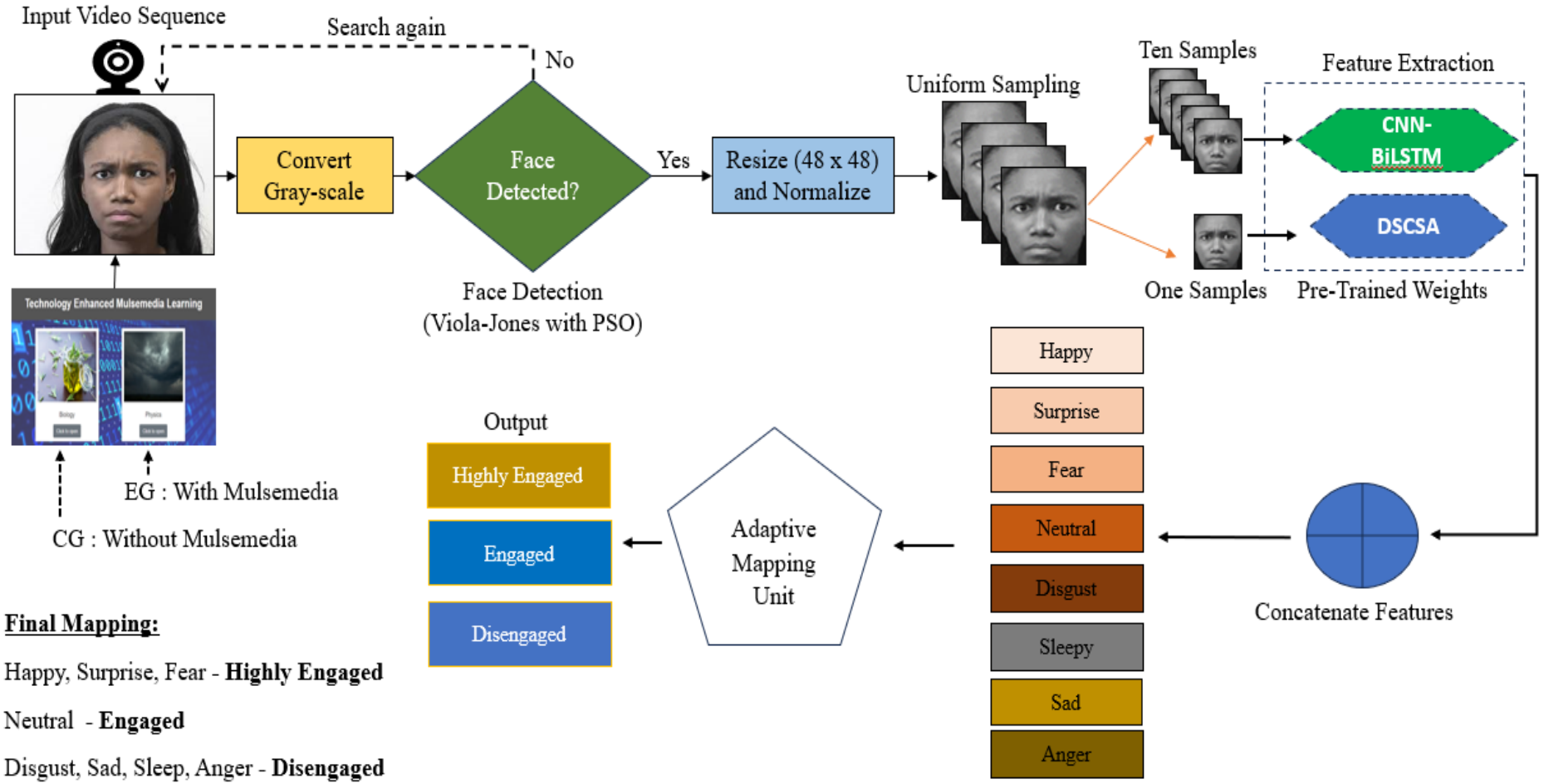

Figure 3.4 Analysing Learner's Engagement in mulsemedia-synchronised Learning Environment

3.4 Chapter Summary

This chapter briefly discusses the proposed research methodology and outlines the process of detecting learner engagement through FER in a mulsemedia learning environment. It includes face detection with pre-processing techniques, feature extraction methods, emotion recognition, and mapping universal facial expressions to determine learners' engagement levels. The research makes four key contributions: first, the prediction accuracy of the conventional Viola-Jones face detection algorithm is improved using the PSO technique. Second, a CNN-BiLSTM architecture is proposed to extract and analyze spatiotemporal features from a sequence of images using a supervised learning approach. Third, a semi-supervised convolutional sparse autoencoder approach enhances FER prediction accuracy with fewer samples, addressing overfitting issues in imbalanced datasets. Fourth, an immersive learning approach based on mulsemedia is introduced to assess learner satisfaction through a QoE questionnaire. Additionally, the FER approach is utilized to analyze learner engagement both with and without mulsemedia. The results indicate that the mulsemedia-synchronized learning environment significantly improves memory retention and attention to learning content compared to conventional learning methods.