■ Deep Learning Project Exam

Topic: Video Emotion Recognition using CNN, LSTM, and Autoencoder

Objective:

Design and implement a deep learning model that can recognize human emotions from short video clips by combining Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and Autoencoders.

Learning Outcomes:

1. Understand CNNs for spatial feature extraction.
2. Understand LSTMs for temporal sequence modeling.
3. Understand Autoencoders for unsupervised feature learning.
4. Combine multiple architectures in a single deep learning pipeline.
5. Evaluate performance using accuracy, F1-score, and confusion matrix.

Dataset:

You can use the CREMA-D dataset available on Kaggle (https://www.kaggle.com/datasets/ejlok1/cremad) or a subset of FER2013 dataset transformed into video sequences. Each video contains 10–20 frames representing different emotional expressions.

Tasks to Complete:

1. 1. Extract frames from each video and preprocess them (resize, normalize, grayscale).
2. 2. Build an Autoencoder to compress and denoise frame features.
3. 3. Build a CNN model to extract spatial features from each frame (ResNet or custom CNN).
4. 4. Feed the CNN-encoded frame features into an LSTM to capture temporal dependencies.
5. 5. Add fully connected layers and a Softmax output for emotion classification.
6. 6. Train, validate, and test your model. Record loss and accuracy curves.
7. 7. Evaluate your model using a confusion matrix, precision, recall, and F1-score.
8. 8. Write a short 1–2 page report explaining your design choices, results, and improvements.

Suggested Architecture:

Input (T × H × W × C) → CNN Encoder → Autoencoder Bottleneck → LSTM → Dense → Softmax

Evaluation Metrics:

1. Accuracy
2. Precision, Recall, F1-score

3. Confusion matrix
4. Optional: t-SNE or PCA visualization of features

Bonus (for advanced students):

1. Add an attention mechanism on LSTM outputs.
2. Replace CNN + LSTM with a 3D CNN for spatiotemporal processing.
3. Compare Autoencoder+CNN-LSTM vs CNN-LSTM performances.

Grading Rubric:

1. Data preprocessing & frame extraction – 10 pts
2. Model architecture correctness (CNN, LSTM, Autoencoder) – 30 pts
3. Training & tuning quality – 20 pts
4. Evaluation & metrics interpretation – 20 pts
5. Report clarity & reproducibility – 20 pts