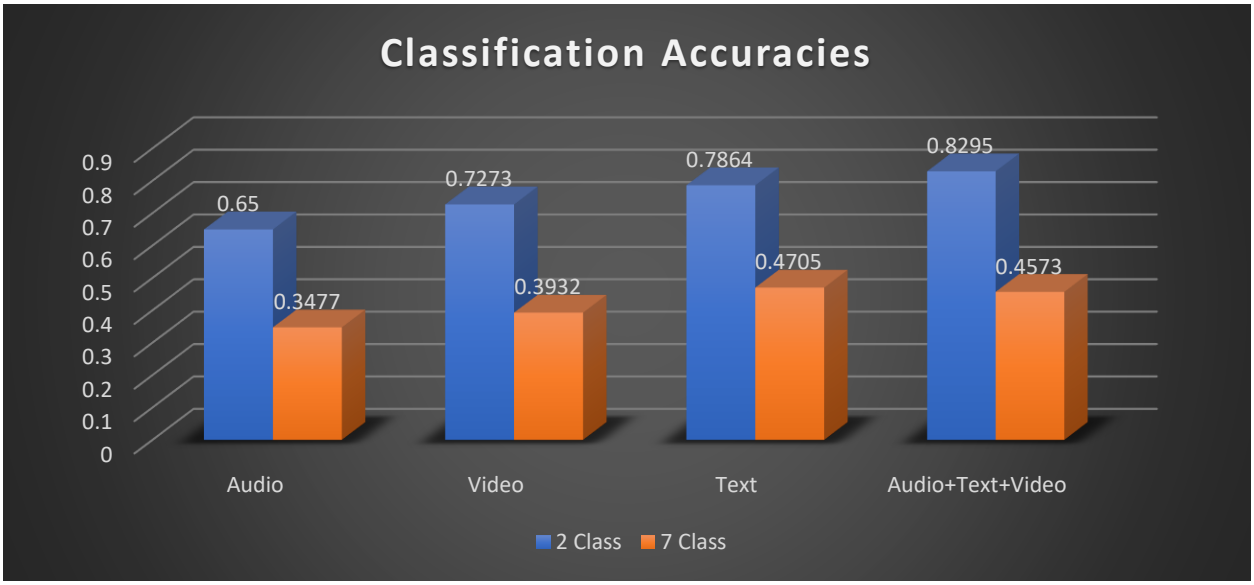
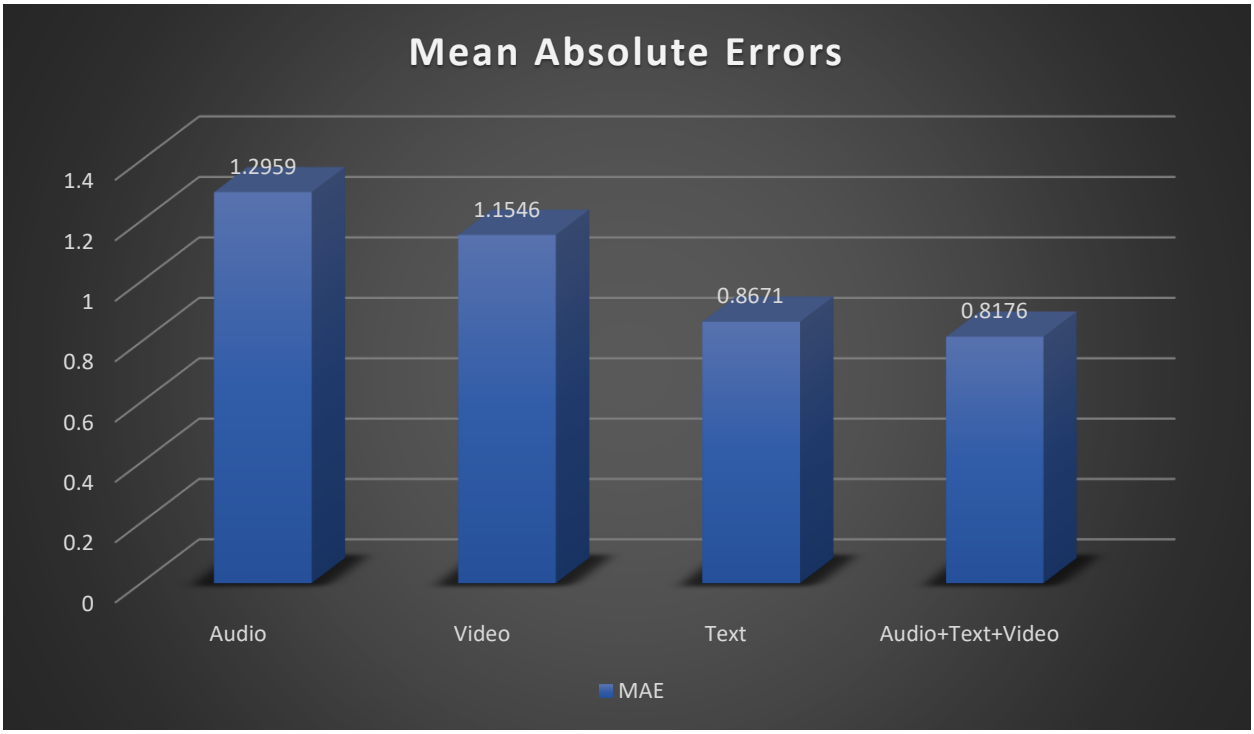
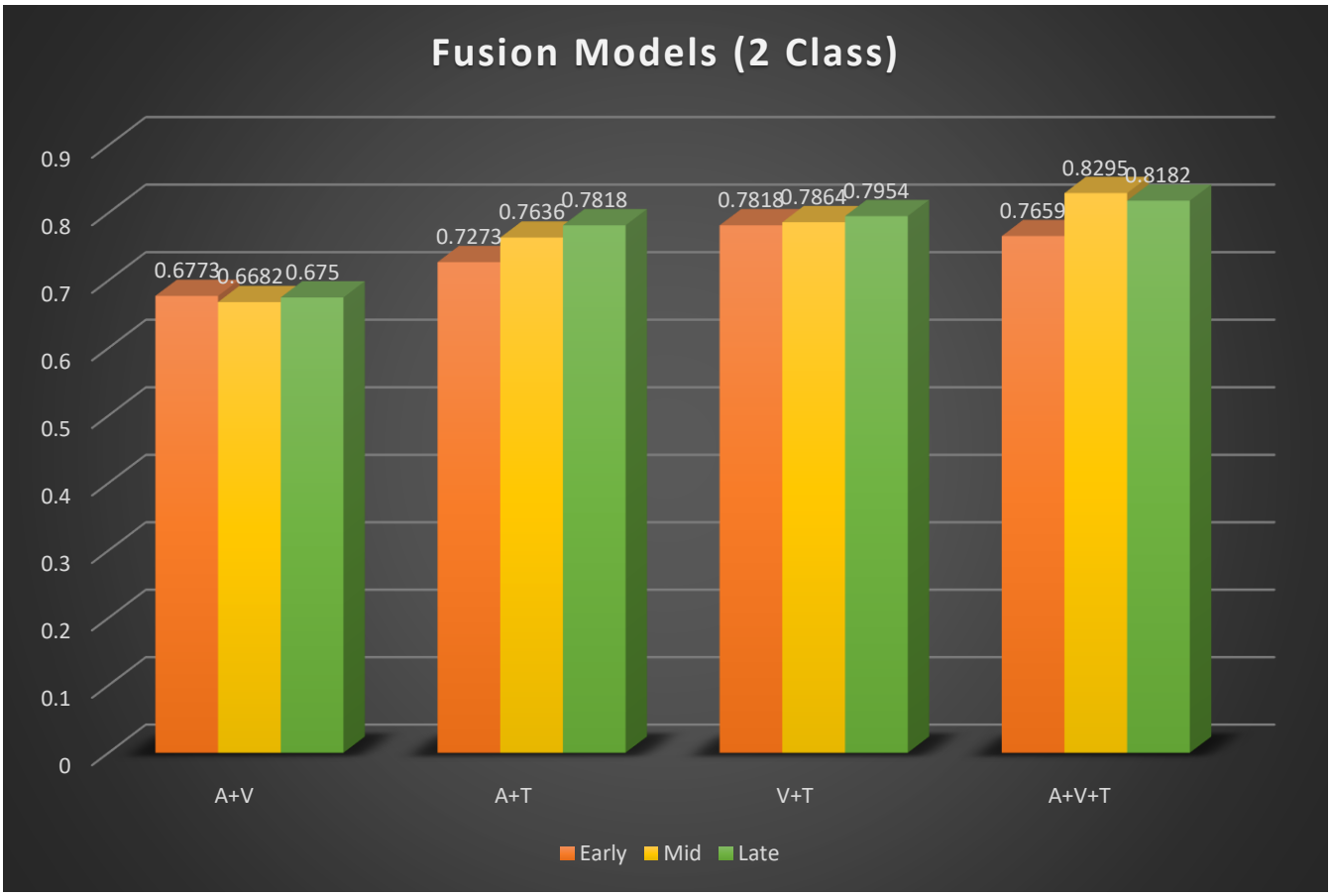
Train: Val: Test Split -> 70:10:20





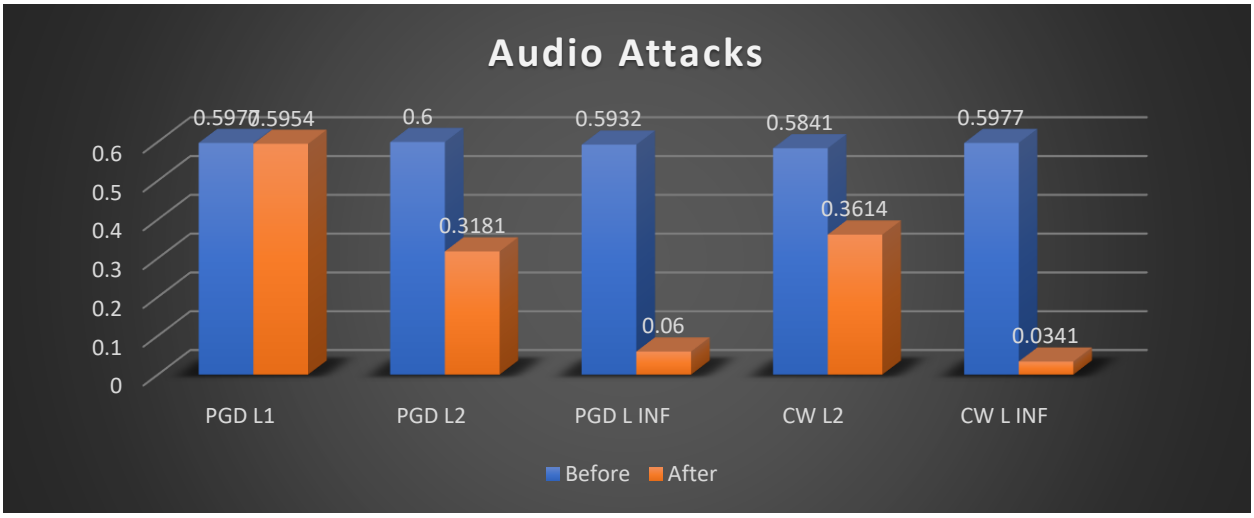


Early Fusion: Concatenation is done at the input layer

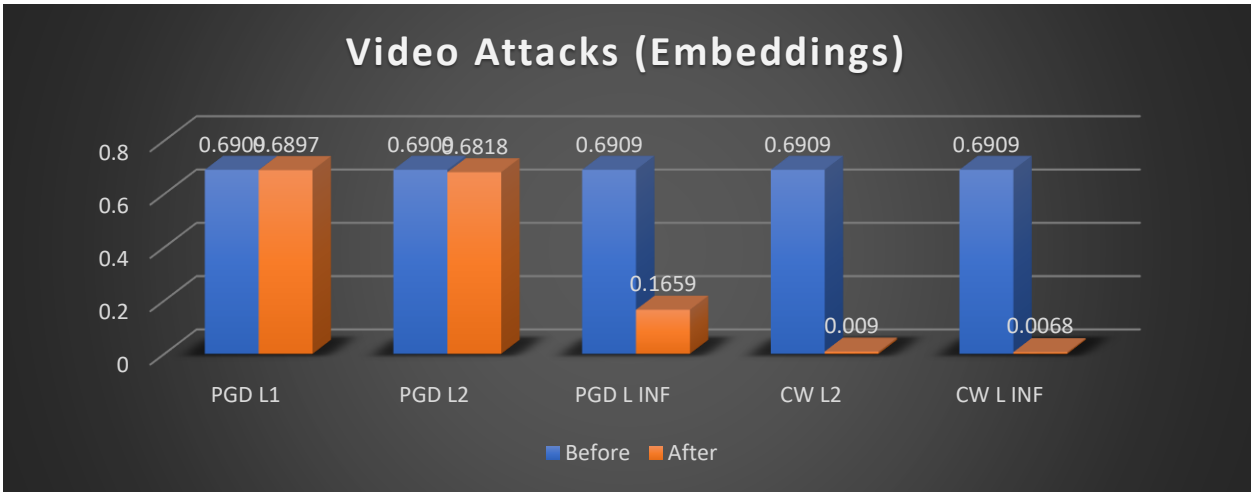
Mid Fusion: Output of unimodal Transformer models are concatenated, passed through layer normalization, two linear layers and the output layer

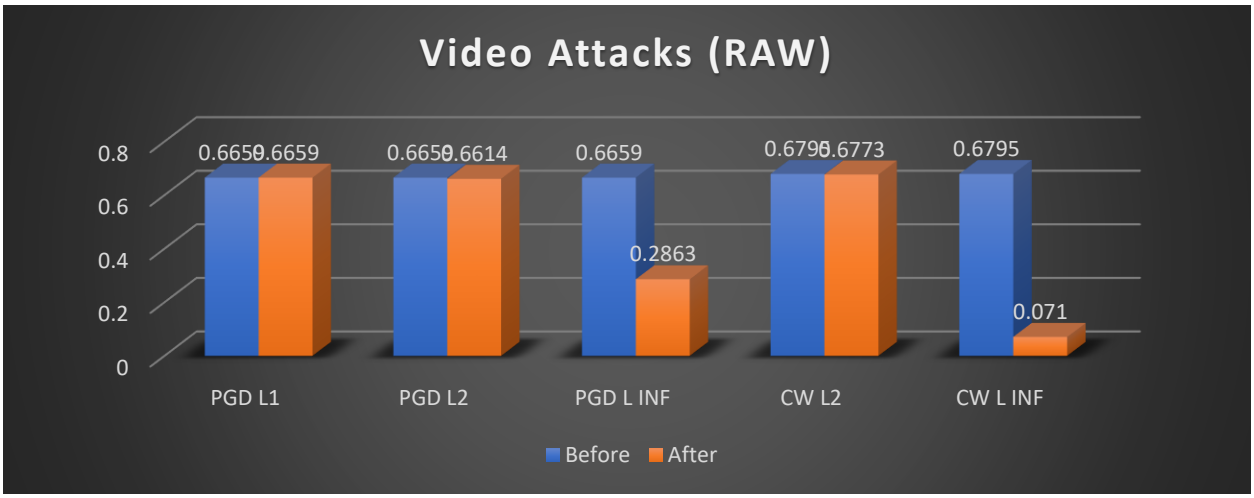
Late Fusion: Weighted sum of logits of the individual models are taken and passed through a SoftMax layer. For A+V+T, weights are 0.2, 0.3 and 0.5, respectively. For A+T model, weights are 0.25 and 0.75. For V+T, weights are 0.35 and 0.65. For A+V, weights are 0.4 and 0.6.

2 CLASS Attacks

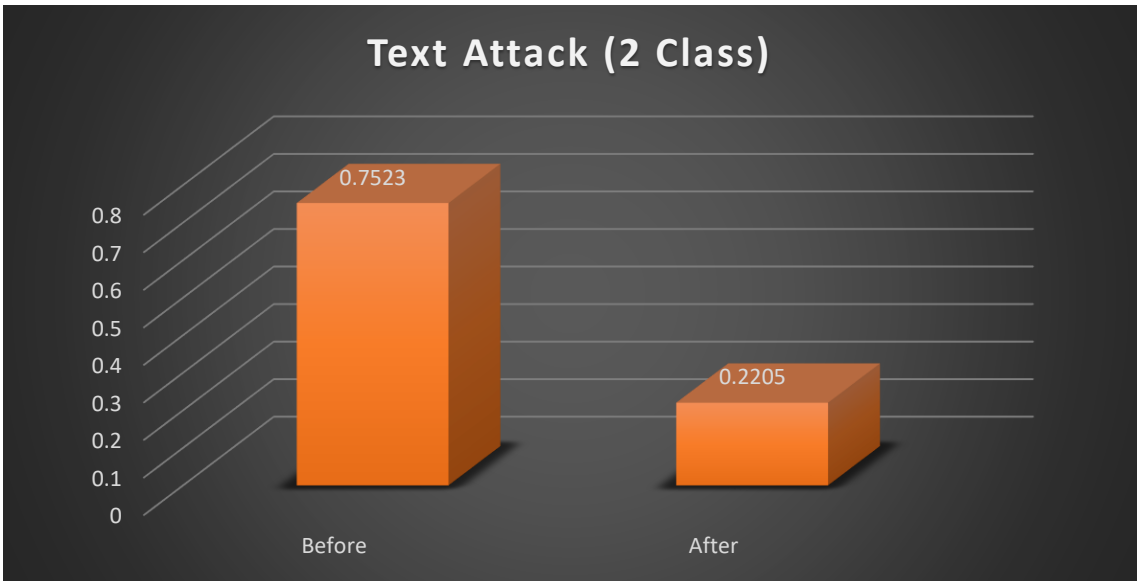


Attacks are performed on Audio model only.

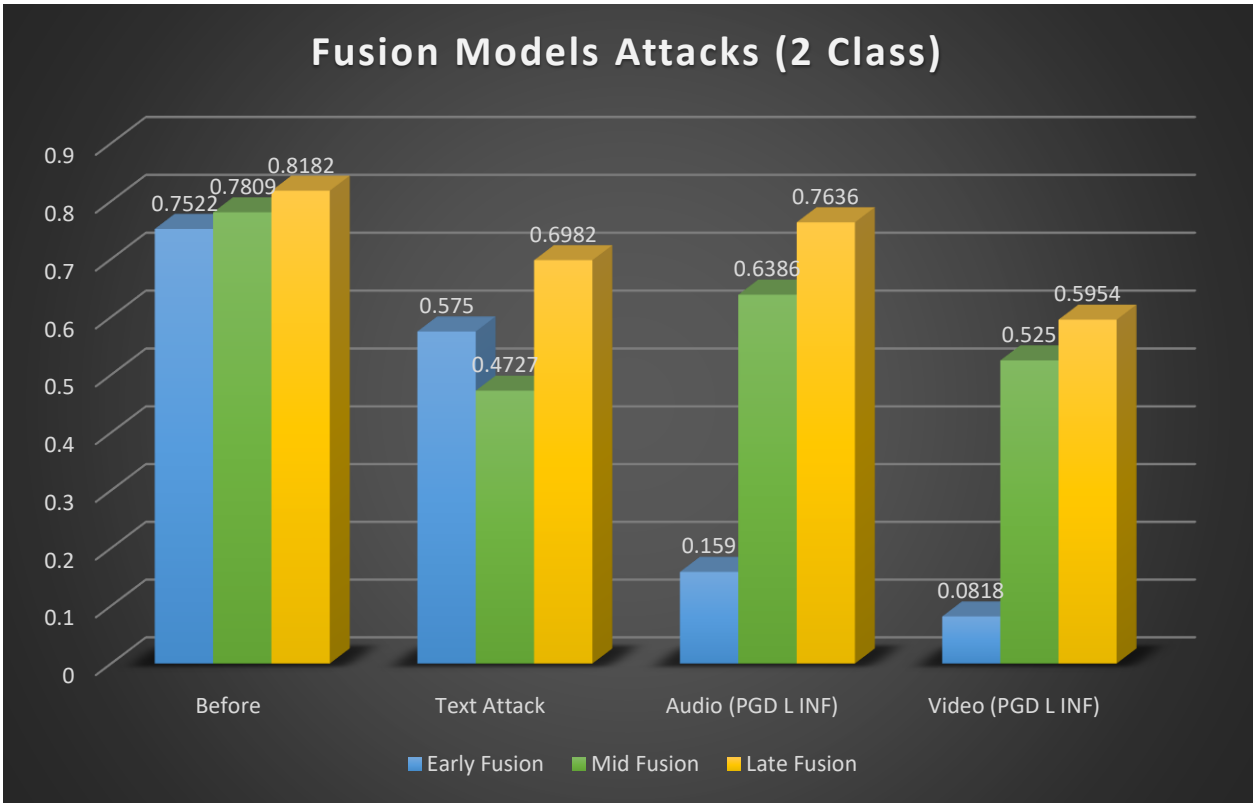




Attacks are performed on video models only.



Attack is performed on Text model only.



Only 1 attack is used at a time. Attacks are generated using Individual models. That is, for text attack, adversarial reviews were generated using a Text only model