Group members:

Sohan kumar-22052159

Abhinav -2205870

Srimats -22052162

AYUSH KUMAR:22052107

Deepfake Detection Using Explainable AI (XAI)

The rapid advancement of deepfake technology poses significant challenges to digital media integrity, with potential threats spanning from misinformation to identity fraud.

While deep learning models have demonstrated remarkable accuracy in detecting deepfakes, their inherent black-box nature raises concerns regarding trust, interpretability, and real-world deployment.

In this study, we explore the integration of Explainable AI (XAI) techniques to enhance the transparency and reliability of deepfake detection models.

By leveraging interpretability methods such as SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and Grad-CAM (Gradient-weighted Class Activation Mapping),

we aim to provide insights into model decision-making processes.

Our approach evaluates the performance of popular deepfake detection models, including CNN-based and transformer-based architectures, across benchmark datasets such as FaceForensics++ and Celeb-DF.

The results of this study contribute to the development of more trustworthy and accountable AI systems for deepfake detection, ensuring robustness against adversarial manipulations and improving user confidence in AI-driven solutions.

Future work will focus on optimizing the explainability-performance trade-off and deploying interpretable models in real-world applications.