UNIVERSITY OF WESTMINSTER

INFORMATICS INSTITUTE OF TECHNOLOGY

INFORMATICS INSTITUTE OF TECHNOLOGY
In Collaboration with
UNIVERSITY OF WESTMINSTER, LONDON

XAIVIER: A Visually Interpretable Forensic Deepfake Detection Tool Using Anchors

A Dissertation by
Krishnakripa Jayakumar

Supervised by
Dr. Nimalaprakasan Skandhakumar

Submitted in partial fulfilment of the requirements for the
MSc in Cyber Security and Forensics degree
at the University of Westminster.

May 2022

Abstract

Cyber-crimes such as manipulations of videos that threaten to violate the privacy and identity of a person by hijacking their face and swapping it onto the body of another person (face-swapping) has seen a dramatic rise over the recent years. The "Deepfakes" of today, as they are known, have become very convincing that they are indistinguishable from authentic videos. Though promising headway has been made in field of deepfake detection, existing research mostly focus only on classifying a video in binary forms- either as a deepfake or not, without any explanations as to why the model classified it as such. However, these works fail in situations where explainability and transparency behind a tool's decision is crucial, especially in a court of law, where the Court may demand justifications and explanations for why a video is a deepfake from a digital forensic expert. Explainable AI (XAI) has the potential to give a whole new meaning to deepfake detection, hence a new research area called Explainable Deepfake detection comes into the picture that help explain decisions behind a black-box system's predictions, easily to humans. This is however a new yet niche area with limited contribution, therefore, based on the limitations identified in the existing XDD works, this research proposes the use of an XAI method called Anchors, a model-agnostic high precision explainer, that can explain the reasons behind the predictions of a custom made Deepfake detector. This work is also one of the first to approach the problem from a digital forensic viewpoint, enabling the development of a digital forensic toolkit for deepfake detection. The approach, design, implementation, and evaluation steps are detailed in this report.

Subject Descriptors: Deepfake Detection, Explainable AI (XAI), Interpretable Machine Learning (IML), Digital Media Forensics

Keywords: Deepfake, XAI, IML, DCNN, Computer Vision, Efficient-Net, Digital Media Forensics