Revised Project Proposal

A deepfake is an artificial video or image where a person's face in said video or image is replaced with another person's face or is manipulated in some way. As of now, deepfakes are incredibly realistic and it is difficult for an average person to distinguish between a deepfaked video and an authetnic one. Therefore, for my final project, I would like to investigate the topic of deepfake detection using machine learning. Since deepfakes are so realistic, machine learning has been a viable way to detect deepfaked images or videos.

There are three types of deepfakes: face synthesis, face swap, and ~~attribute~~ expression modification. Face synthesis involves creating completely artificial faces using provided data. Examples of such images can be found here: https://this-person-does-not-exist.com/en. Although the faces on this website look realistic, they are not real. Face swaps involve swapping the face of a person in an image or video with another person's face. ~~Attribute~~ Expression modification involves modifying the ~~attributes~~ expressions of a person's face (such as whether the person is smiling or not, eye color, etc.). This is usually done by transferring the expressions of a source actor to a target actor while preserving the target actor's identity. Unlike face swaps, the original person is still there but has a different expression. It can even involve changing the age of a person in a photo or video. Given these three categories of deepfakes, I want to investigate how strong or weak different models of deepfake detection models/classifiers are for the three categories. I will be using ~~two~~ one popular dataset: ~~DFDC (https://www.kaggle.com/competitions/deepfake-detection-challenge/data) and~~ faceforensics++ (https://www.kaggle.com/datasets/sorokin/faceforensics). Additionally, I will be using MesoNet (https://github.com/DariusAf/MesoNet) and Deepfake scanner (https://github.com/deepware/deepfake-scanner), as classifiers as these two are very reliable. If time permits, I will also use https://github.com/aaronchong888/DeepFake-Detect. ~~others (https://github.com/aaronchong888/DeepFake-Detect, https://github.com/osalpekar/DeepFake-Detection, https://pythonlang.dev/repo/dessa-oss-deepfake-detection/, https://pythonlang.dev/repo/bbvanexttechnologies-fakevideoforensics/) as classifiers.~~ I will be comparing the results of the classifiers I use to each other.

Additionally, I plan to see how modifying different attributes of a person's face affects the results of deepfake detectors (This part of my project will only involve attribute modification). For example, does changing a person's hair color make it easier for a deep fake detector to detect a deepfake than changing the person's eye color? This part of my project may involve generating my own deepfakes with custom attributes. I plan to do this using styleGAN ~~or Deepfacelab~~. These tools will allow me to create my own deep fake images.

This project is relevant to the course because it is about a recent and ongoing issue in software engineering. The papers I have read so far were published in 2019 or after. I chose this topic because I find the fact that machine learning models can generate artificial yet realistic images very interesting. By doing this project, I get to expand on my interest by investigating how machine learning can distinguish between what is real and what is fake. Additionally, I believe that my work can help researchers because my data will show them how they can

improve on existing models. Since my project investigates the strengths and weaknesses of current deepfake detectors, researchers may get a better understanding of how to improve existing models.