UNIVERSIDAD YACHAY TECH

ESCUELA DE CIENCIAS FÍSICAS Y NANOTECNOLOGÍA

"Review Paper on Algorithms and Advancements in Detecting Deepfakes"

Ángel Salazar, Johel Cuasapaz, Iván Fierro
(angel,salazar; johel.cuasapaz; ivan.fierro)@yachaytech.edu.ec

Abstract

The proliferation of AI-generated synthetic media (deepfakes) has increased misinformation on social media. This review evaluates recent automated deep learning models for deepfake detection. It analyzes how architectures like Convolutional Neural Networks (CNNs), Autoencoders, Generative Adversarial Networks (GANs), and Diffusion Models distinguish deepfakes from authentic content in real time. The review emphasizes comparative accuracy on manipulated images and audiovisual content, and addresses ethical considerations and future directions for these detection systems.

Keywords: deepfake detection, machine learning, CNNs, GANs, synthetic media, artificial intelligence.

Introduction

The exponential growth and and popularization of generative AI, has encourage synthetic content known as deepfakes. Which are modified images and videos created with misinformation and fraud objectives.

The current problem is that:

- Deepfake files have 1500% increase in 2023 [1]
- Deepfake-related fraud increase in 3000% in 2023 [1]
- The human detection rate for high-quality manipulated videos remains critically low, at approximately 24.5%.[2]

This failure of human detection underscores the urgent need for detection systems. For that reason, We move toward machine learning (ML) and deep learning (DL) models, where DL models are more powerful, learning patterns from large datasets.
Nowadays the Deppfake detection can be classified into: Feature-Based Methods, Deep Learning-Based Methods, Hybrid Methods (Combining both)
This review focuses on the second and third types, analyzing their effectiveness with real-world data according to current scientific research.

State of the art

There is principally 2 methods to detect Deepfakes:

- Machine Learning Methods: these methos use a tree-based approach for claisification, susch as Support vector Machine (SVM) that learns by example to assign labels to objects, Logistic Rregrssion (LR) that applies logit transfotmation to probabilities, and Multilayer Perceptron (MLP) that is a neural network that uses activation functions to model complex relations.
- Deep Learning Methods: there is some of them such as Convolutional Neural Networks (CNNs) that is designed to process spatial dara, Recurrent Neural Networks (RNNs) that is used to detect patterns in dara sequences, and GAN Simulators.

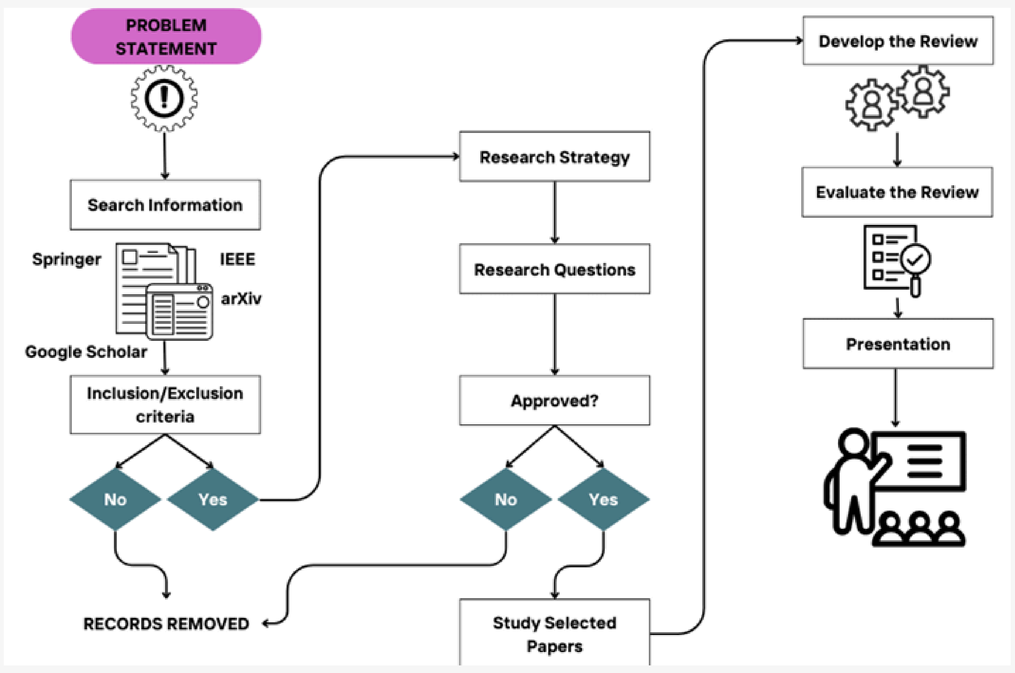
Large-scale datasets are primary tools for training deepfake detectors. However, variations in image quality can cause problems in sophisticated environments [3]. Current detection can be divided ino three categories: Navive Detector, Spatial Detector and Frequency Detector.

Discussion and conclusions.

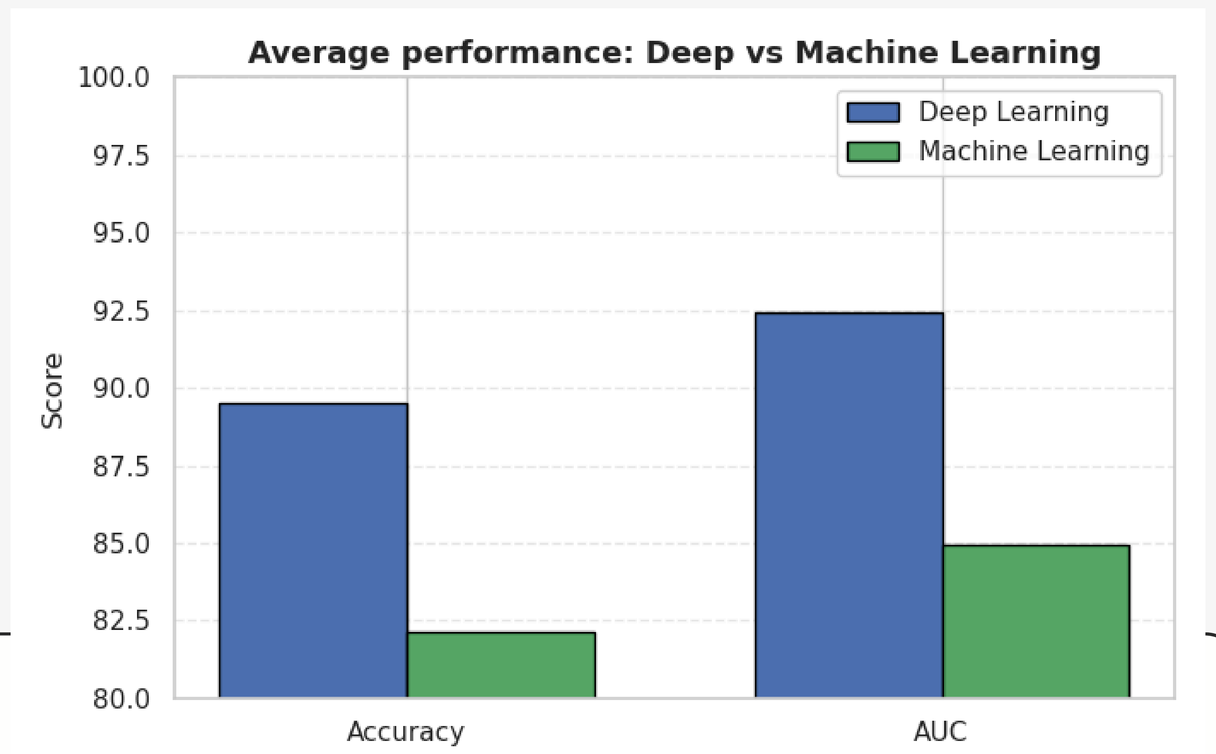
Deep Learning models are better than Machine Learning detecting Deepfakes, often exceeding 95% accuracy. Howevern the performance depends on the size and quality of the datasets. There ir High accuracy (~98%) on simpler datasets like FaceForensics++. On complex, real-world fatasets like DFDC, there is a significant drop (to ~64%).

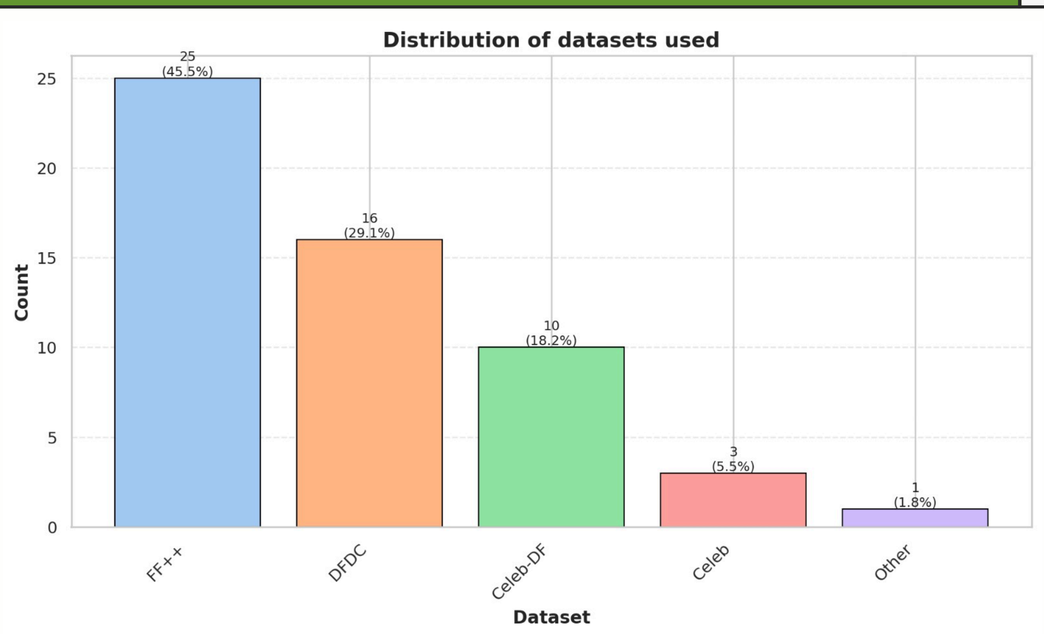
As a conclusion. Deep learning is the most effective approach for Deepfake detection.
The power of current models needs improvement to handle diverse real-world content, as this is vast and complex, revealing a generalization problem when trying to detect deepfakes.
Ethical issues such as privacy and false positives require urgent attention to ensure that these tools, in addition to being appropriate, reliable, and effective, respect public privacy and security.

Materials and methods



Results





For deep learning, most points are clustered in the upper region of the graphs, which confirms that high performance is not driven by a single architecture but rather represents a consistent pattern across different network families. Nonetheless, there are also a few outliers with much lower accuracy on FF++ (e.g., some early Xception and RNN variants).

On the machine learning side, the scores are concentrated in a noticeably lower band, with no model reaching the best-performing deep networks in either accuracy or AUC.

Bibliographic references

[1] DeepStrike. Deepfake Statistics 2025. DeepStrike 2025. Accessed: 2024-11-20.
[2] Köbis, N.C.; Doležalová, B.; Soraperra, I. Fooled twice: People cannot detect deepfakes but 321 think they can. Iscience 2021, 24 [
[3] Guarnera, L.; Giudice, O.; Battiato, S. Fighting deepfake by exposing the convolutional traces 369 on images. IEEE access 2020, 8, 165085–165098.