IBIMA Publishing
Communications of the IBIMA
https://ibimapublishing.com/articles/CIBIMA/2024/457767/
Vol. 2024 (2024), Article ID 457767, 9 pages, ISSN: 1943-7765
https://doi.org/10.5171/2024.457767

Research Article

Performance Analysis on Deep Fake Detection

[1]Stéphane MONTEIRO, [2]Cristina WANZELLER and [2]Filipe CALDEIRA

[1]Polytechnic of Viseu, Viseu, Portugal

[2]CISeD – Research Centre in Digital Services, Polytechnic of Viseu, Viseu, Portugal

Correspondence should be addressed to: Stéphane MONTEIRO; estgv16842@alunos.estgv.ipv.pt

Received date:14 March 2024; Accepted date:11 July 2024; Published date: 3 September 2024

Academic Editor: Luís Cavique

Abstract

A deepfake is a type of synthetic media, image, video or audio of a person in which their physiognomy has been digitally altered, using artificial intelligence, particularly deep learning techniques, so that they appear to be someone else, typically used maliciously or to spread false information. In this study, our main goal is to thoroughly assess the effectiveness of deepfake detection algorithms by using various key performance metrics such as accuracy, precision, recall, and F1-score. The primary focus of our analysis revolves around their capability to distinguish between genuine and manipulated videos.

Furthermore, our research involves a detailed examination of specific types of deepfake manipulations with the aim of identifying differences in detection accuracy and performance across these categories. We go beyond just analyzing the algorithms and investigate how characteristics of the dataset, like diversity and size, impact the detection performance of the tested algorithms.

We anticipate that the results of this research will make substantial contributions to the advancement of deepfake detection technology. Furthermore, the insights obtained from this study will not only assist in refining existing detection algorithms but also offer valuable guidance for future research in the field of deepfake detection, ultimately contributing to the ongoing fight against the spread of deceptive digital media.

Keywords: Deepfakes, Machine Learning, Deep learning, DFDC, XCeption, ResNet, VGG

Cite this Article as: Stéphane MONTEIRO, Cristina WANZELLER and Filipe CALDEIRA (2024)," Performance Analysis on Deep Fake Detection ", Communications of the IBIMA, Vol. 2024 (2024), Article ID 457767, https://doi.org/10.5171/2024.457767

Introduction

Deepfake videos have become a significant issue in today's digital landscape, as the rapid development of deep learning techniques has made it increasingly challenging to distinguish between real and fake videos. This has raised concerns about the potential misuse of this technology. Developing effective deepfake detection algorithms is crucial in addressing this pressing problem.

The DeepFake Detection Challenge (DFDC) dataset has emerged as a benchmark dataset for evaluating and enhancing deepfake detection techniques. This dataset includes a variety of deepfake and authentic videos that utilize different visual editing methods, covering a wide range of topics and scenarios. For researchers, the DFDC dataset is an invaluable resource for the development and evaluation of deepfake detection algorithms (Ghazi and Ekenel, 2016).

In this paper, a comprehensive performance analysis of three widely utilized deepfake detection algorithms - XCeption, ResNet, and VGG - is presented. The aim is to gain insights into their effectiveness and identify areas that may require further development by assessing their performance on the DFDC dataset.

The research evaluates the accuracy, precision, recall, and F1-score of the XCeption, ResNet, and VGG algorithms in distinguishing between authentic and deepfake videos. General performance across the dataset and specific performance for various subsets based on modification methods, video quality, and facial characteristics are considered. This approach enables a better understanding of the strengths and weaknesses of these algorithms across various variables.

The findings of this performance analysis are expected to drive advancements in deepfake detection technology. Researchers and practitioners striving to tackle the challenges posed by deepfake videos will find the insights gained from evaluating the XCeption, ResNet, and VGG algorithms on the DFDC dataset highly valuable. The research will also highlight areas for further exploration, algorithm enhancement, and the development of more reliable deepfake detection techniques.

Throughout this study, the performance of the XCeption, ResNet, and VGG detection algorithms on the DFDC dataset is thoroughly analyzed. The assessment results will provide crucial information and guidance for ongoing research and development in deepfake detection by demonstrating the efficacy of these algorithms in discriminating between real and deepfake videos.

Related Work

Deepfake videos pose a significant challenge for media forensics, particularly in proving video authenticity. There is an urgent need to create robust deepfake video detection algorithms aiming to mitigate potential risks while contributing to minimizing the spread of fake information. This area is getting more attention from the research community, which has led to recent advances in deepfake detection techniques, including various machine learning-based approaches. The following literature review presents some relevant state-of-the-art deepfake detection algorithms detailing their strengths and weaknesses, focusing mainly on VGG, ResNet, and Xception.

1. Visual Geometry Group (VGG) Architecture

The Visual Geometry Group developed the Visual Geometry Group (VGG) architecture at the University of Oxford. It is a convolutional neural network (CNN) with proven performance for visual recognition. VGG can be exploited for deepfake detection feature extraction because it can capture detailed spatial hierarchies in images. It also contributes to identifying artifacts and irregularities introduced by deepfake generation techniques. Deep convolutional layers refer to a type of layer used in deep learning models, particularly convolutional neural networks (CNNs), designed for processing structured grid data, such as images. The deep convolutional layers in the VGG architecture have been extensively employed for deepfake detection. VGG models that have already been trained are used in methods like VGGFace (Ghazi and Ekenel, 2016) to extract high-level face characteristics and spot discrepancies brought on by deepfake manipulations (Chang *et al.*, 2020).

2. Residual Neural Network (ResNet)

Residual Neural Networks (ResNets) are a variety of deep neural networks aiming to minimize the vanishing gradient problem, allowing the training of very deep networks. Deepfake detection challenges were successful for ResNet's residual learning architecture (Yadav, no date). For increased detection accuracy, techniques like

ResNet-based Fusion (Pashine *et al.*, 2021) use ensemble models integrating ResNet-50 and ResNet-101 (the number indicates the number of layers), to capture global and local information.

3. XCeption

An expansion of the Inception architecture called Xception (Extreme Inception) has proven to perform especially well in several computer vision applications, including deepfake detection. Techniques such as Xception-DF, an adaptation of the Xception model designed explicitly for deepfake detection (Atas, Ilhan and Karakse, 2022), use Xception networks to extract discriminative characteristics from modified face areas and spot irregularities. Xception-DF leverages the architecture of Xception by identifying subtle operations in video content with high precision.

4. Ensemble Approaches

Deepfake detection was leveraged by ensemble approaches combining various models or algorithms. To improve detection accuracy, fusion-based ensembles, such as VGG16 + Xception, combine the predictions of the VGG16 and Xception models (Khatri, Borar and Garg, 2023).

Despite the advancements achieved, deepfake detection systems still face multiple challenges. On the one hand, existing deepfake methods constantly evolve, posing a threat that necessitates algorithm adaptation to new manipulations. On the other hand, the detection models are generally constrained due to a lack of extensive and varied datasets, emphasizing the demand for additional representative datasets.

The DeepFake Detection Challenge (DFDC) dataset was developed in collaboration with Facebook, Inc. and Microsoft to address the increasing concerns related to the spread of deepfake videos (Trabelsi, Pic and Dugelay, 2022). Leveraging artificial intelligence algorithms to control facial expressions, body movements, and voice, actual deepfakes may produce exceptionally realistic fake videos. With the constant technological advances, the need for robust deepfake detection algorithms grows and becomes critical in combating disinformation, which may cause threats to privacy and decrease faith in the media. The DFDC datasets aim to be a consistent baseline for assessing and improving deepfake detection techniques. They allow researchers and developers to train, test, and evaluate their deepfake detection algorithms in a controlled and consistent approach by providing a wide selection of authentic and deepfake videos (Dolhansky *et al.*, 2020). The dataset comprises diversified content, including a wide range of manipulation techniques and people, ensuring that the generated algorithms are durable and capable of generalizing to real-world deepfake scenarios.

Moreover, the DFDC dataset supports study collaboration by providing a venue for academics to compare detection algorithms, discuss ideas, and collaboratively address the issues supplied by deepfake technology. The dataset's availability on platforms like Kaggle improves information exchange and successfully stimulates novel solutions to address deepfake threats (Dolhansky *et al.*, 2019). In conclusion, the development and availability of the DFDC dataset represent an essential step in encouraging the development of accurate and effective deepfake detection algorithms, contributing to the ongoing efforts to limit the potential threats posed by deepfake videos in today's digital ecosystem.

Methodology

This section outlines the methodology employed in our study, encompassing data preprocessing techniques, model architecture specifications, training protocols, and evaluation metrics.

1. Data Preprocessing

One essential step is preprocessing data. It allows the preparation of the DFDC dataset for deepfake detection. The main objective is to convert raw video data into an appropriate format for training deep learning models. The data preparation process applied in this work is as follows:

- **Frame Extraction:** Individual frames were obtained from each video in the DFDC dataset. For consistency across the dataset, this technique required sampling frames at a particular frame rate (for example, 30 frames per second). The previous steps allowed the split up of the video clips into several groups of distinct frames.
- **Data Splitting:** The DFDC dataset was divided into training, validation, and test sets. A substantial portion was allocated for training, a smaller segment for validation to monitor progress and adjust hyperparameters, and the final portion for an unbiased evaluation of model performance.

The above preprocessing steps were essential for transforming raw video data into a suitable format for training deep learning models. This meticulous preprocessing workflow subjected our deepfake detection models to diverse variables such as facial expressions, lighting conditions, and video quality. Consequently, they have become more accurate at identifying deepfake content within the DFDC.

2. *Model Architectures*

In our study on deepfake detection, we utilized three different deep learning architectures, each possessing distinct characteristics and benefits.

XCeption

The XCeption architecture comprises a deep convolutional neural network (CNN) intended to detect complex features available in images while keeping the model weight under control (Rismiyati *et al.*, 2020). Based on experience, XCeption proves highly suitable for tasks such as deepfake detection and other image-related tasks. A thorough summary is presented:

- **Architecture:** XCeption distinguishes itself through the use of depth-wise separable convolutions in its inception-style modules, which enables the network to capture features across different scales efficiently with fewer parameters (Pan *et al.*, 2020). It represents an enhanced iteration of the Inception architecture, incorporating advancements in convolutional operations.
- **Transfer Learning:** The core of our deepfake detection was an XCeption model that had already been trained. This signifies that we trained the model's weights on a sizable dataset before initializing it. Before fine-tuning the DFDC dataset, pre-training enables the model to pick up essential characteristics from various images.
- **Fine-tuning:** After initializing the XCeption model with pre-trained weights, we refined its performance on the DFDC dataset. This fine-tuning process involves adapting the model specifically for deepfake detection. The use of pre-trained weights provides a solid foundation while fine-tuning tailors the model to meet the specific demands of the task.

ResNet

ResNet, also known as Residual Neural Network, is renowned for its residual learning blocks, enabling effective training of exceptionally deep neural networks. (Rismiyati *et al.*, 2020).

- **Architecture:** ResNet incorporates residual blocks with skip connections throughout its architecture. These skip connections facilitate smoother gradient flow, particularly in very deep networks. The ResNet-50 and ResNet-101 variants denote the number of residual blocks used in these architectures. (Mukti and Biswas, 2019).
- **Ensemble Approach:** Rather than selecting a single ResNet design, we adopted an ensemble approach by combining ResNet-50 and ResNet-101. Ensemble learning involves integrating predictions from multiple models to enhance the overall performance. By leveraging both ResNet-50 and ResNet-101, our aim is to effectively capture both local and global properties.
- **Transfer Learning and Fine-Tuning:** Like XCeption, we initialized the ResNet models with pre-trained weights from a large-scale image classification dataset. These pre-trained weights establish a strong starting point for the models. Subsequently, we fine-tuned these models on the DFDC dataset to tailor them specifically for the deepfake detection job.

VGG

The VGG (Visual Geometry Group) architecture is renowned for its simplicity and effectiveness (Rismiyati *et al.*, 2020). We selected the VGG16 architecture based on its established success across diverse computer vision tasks:

- **Architecture:** VGG networks are distinguished by their simple yet effective design, featuring stacked layers and compact 3x3 filters. The VGG16 model, with its 16 layers, is well-regarded for its high performance in numerous computer vision applications. (Tammina, 2019).
- **Pre-trained model:** We began with the pre-trained VGG16 model, leveraging its pre-existing weights to incorporate essential feature representations crucial for precise deepfake detection.
- **Transfer Learning and Fine-Tuning:** Following pretraining, we fine-tuned the VGG16 model using the DFDC dataset. This process allows the model to adjust to the nuances and challenges inherent in deepfake detection, optimizing its performance for this specific task.

3. *Evaluation Metrics*

In our deepfake detection research, we evaluated the effectiveness of the XCeption, ResNet, and VGG algorithms using commonly employed evaluation metrics listed below.

In evaluating deepfake detection systems, two crucial metrics take precedence: accuracy and response time. Accuracy serves as the cornerstone of reliability, measuring the system's ability to distinguish genuine content from manipulated material and thereby minimizing false positives and false negatives. Meanwhile, response time is equally critical in determining the algorithm's practicality, especially in time-sensitive applications like video streaming and social media content monitoring. Balancing high accuracy with swift decision-making is essential for advancing deepfake detection systems, ensuring our digital environments remain effective and trustworthy.

- **Accuracy:** Accuracy serves as a pivotal metric for evaluating the performance of each deepfake detection algorithm. It quantifies the percentage of correctly identified samples relative to the entire dataset. In our study, accuracy specifically assesses the algorithms' ability to differentiate between deepfake images and authentic ones. A higher accuracy score indicates greater algorithm efficiency.

$$\text{Accuracy} = \frac{\text{TN+TP}}{\text{TP+FP+TN+FN}}$$

TN – True negative; TP – True positive; FN – False negative; FP – False positive

- **Precision:** Precision, also known as positive predictive value, is a critical metric that reflects the algorithm's ability to minimize false alarms. It measures the percentage of correctly identified deepfakes out of all samples predicted as positive (all samples identified as deepfakes). In the context of deepfake detection, precision evaluates how reliably the system identifies genuine instances of deepfakes. A higher precision indicates a model that produces fewer false alarms.

$$\text{Precision}= \frac{\text{TP}}{\text{TP+FP}}$$

TP – True positive; FP – False positive

- **Recall (Sensitivity):** Recall, also known as the true positive rate, gauges how well each algorithm can reliably identify all deepfake movies. It determines the proportion of accurate positive predictions to all the actual positive examples in the collection. A high recall score means the system successfully detects deepfakes, reducing the likelihood of false negatives.

$$\text{Recall}= \frac{\text{TP}}{\text{TP+FN}}$$

TP – True positive; FN – False negative

- **F1-Score:** The F1-Score, a balanced metric derived from the harmonic mean of recall and precision, provides a comprehensive evaluation of each algorithm's performance. It combines the algorithm's ability to correctly identify all deepfakes (recall) and its precision in distinguishing genuine instances of deepfakes. The F1-Score synthesizes these aspects to gauge how effectively an algorithm can minimize false alarms while accurately detecting deepfakes. Algorithms that achieve a strong balance between accuracy and recall will yield a higher F1-Score.

$$\text{F1Score}= \frac{\text{Precision*Recall}}{\text{Precision+Recall}}$$

DeepFake Detection Challenge Dataset Description

This study presents a comprehensive overview of the DFDC (DeepFake Detection Challenge) dataset, which serves as the foundation for our empirical research. Understanding the characteristics and composition of the dataset is crucial for our investigation. The DFDC dataset is carefully selected, comprising a diverse video collection encompassing various themes and manipulation techniques. This dataset is the training and testing ground for our deepfake detection methods. It includes videos featuring original unaltered content alongside their deepfake counterparts, enabling direct comparisons that facilitate the successful differentiation between authentic and manipulated material. Moreover, the dataset incorporates deepfake techniques like face swapping, facial expression synthesis, and voice modulation. This diversity empowers our detection algorithms to identify a broad range of deepfake content effectively.

In our research, we partitioned the DFDC dataset into three distinct subsets for training, validation, and evaluation of our deepfake detection models (Fig 1). This segmentation is crucial for ensuring the robustness and applicability of our models. Below are the details of this subset division, including the number of images in each subset:

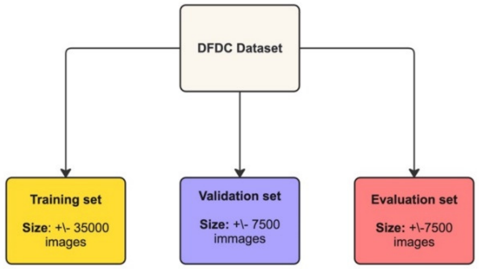

Fig 1. Dataset division

The foundation of our dataset is the training set (Subset 1), which contains about 35,000 images sourced from a variety of videos. Each video in this subset includes both an unaltered original version and a deepfake version featuring different individuals. This training set serves as the basis for our models to learn distinctive features and patterns that distinguish deepfake videos from genuine ones. Importantly, our models utilize transfer learning, initially leveraging pre-trained weights that are then fine-tuned using this training subset.

The validation set (Subset 2) is utilized for fine-tuning and hyperparameter-tuning our deepfake detection models. This subset consists of approximately 7,500 images and serves to monitor the performance and convergence of our models during training. It ensures that our models generalize well beyond the training data, helping to optimize hyperparameters such as learning rates and regularization techniques.

The evaluation set (Subset 3) also consists of approximately 7,500 images. This dataset serves as an uncharted territory to test the effectiveness of our deepfake detection methods. In Subset 3, our models encounter deepfake and authentic images that were not used in training or validation. Assessing our models on this subset offers an impartial evaluation of their ability to generalize to new unseen data.

Performance Analysis

In our experimental setup, we utilized a MacBook Air equipped with the Apple M1 processor, renowned for its exceptional processing power. The hardware configuration included 16 gigabytes (GB) of RAM and a 512-gigabyte (GB) Solid State Drive (SSD) for data storage. We selected this hardware setup based on the M1 chip's proven capabilities in handling AI and machine learning workloads, ensuring smooth execution of our experiments. The 16GB of RAM provided ample memory capacity for rapid data processing and model training. Additionally, the high-speed 512GB SSD facilitated quick data access and storage, which was crucial for managing large datasets and conducting comprehensive testing.

Regarding software, we utilized the Python environment managed by Anaconda version 2021.05, enabling seamless integration of various libraries. TensorFlow 2.5 served as our primary deep learning framework, providing a robust foundation for constructing and training deep neural networks. Additional libraries included Keras 2.4.3 deep learning framework, NumPy 1.19.5 for numerical computations, and Pandas 1.3.3 for efficient data processing. We utilized Jupyter Notebook 6.4.3 as our preferred development environment for its interactive and collaborative features.

At the end of our deepfake detection research, we reach on a crucial phase of our study — the comprehensive evaluation of the performance of our detection algorithms: XCeption, ResNet, and VGG. This phase represents the empirical validation, where we rigorously test these algorithms against the complex challenges presented by deepfake content.

1. Experimental Results

Our initial benchmark comprises a range of metrics. These assessment criteria—accuracy, precision, recall, and the F1-score—thoroughly evaluate the effectiveness of our algorithms in discerning between authentic and deepfake content.

Table 1 provides a comprehensive overview of the performance metrics for the three deepfake detection algorithms—XCeption, ResNet, and VGG.

XCeption achieves an impressive accuracy of 87.5%, making it the top-performing model in the dataset for video classification. VGG follows closely behind with a strong accuracy of 83.7%, demonstrating its effectiveness in classifying videos accurately. In comparison, ResNet achieves an accuracy of 68.7%, performing slightly lower than the other two models. In terms of precision, XCeption demonstrates the highest score at 87.7%, indicating its capability for accurate classifications and minimal false positives. VGG also performs strongly with a precision score of 84.6%, showing a balanced approach between precision and recall. ResNet, with a precision of 69.8%, lags behind, suggesting a comparatively higher rate of false positives.

Table 1 - Evaluation metrics

Evaluation metrics	*Xception*	*Resnet*	*VGG*
Accuracy	87.5%	68.7%	83.7%
Precision	87.7%	69.8%	84.6%
Recall	89.5%	72.3%	89.4%
F1Score	88.6%	71.0%	86.9%
Execution time	87.5ms	280ms	1034ms

In terms of recall, XCeption achieves an impressive 89.5%, indicating its proficiency in identifying a significant portion of actual deepfake content. VGG closely follows with a recall of 89.4%, demonstrating its efficiency in capturing a substantial number of deepfake videos. ResNet achieves a recall of 72.3%, showing effectiveness in recognizing genuine and deepfake content, though it lags behind XCeption and VGG. With a well-balanced trade-off between recall and accuracy, XCeption achieves the highest F1-Score at 88.6%, placing it in the lead. VGG also performs well with an F1-Score of 86.9%, showcasing its ability to accurately differentiate between real and deepfake videos. ResNet exhibits the potential for further optimization, particularly in accuracy, with an F1-Score of 71.0%.

In terms of processing time on the test dataset, XCeption significantly outperformed the other two models. It processed approximately 100ms faster than ResNet and 1000ms faster than VGG.

Conclusion

The advent of deepfake technology has distorted the line between imagination and reality in the ever-evolving digital media landscape. Developing reliable detection algorithms has become increasingly critical as deepfakes gain popularity. This research aims to evaluate three prominent deepfake detection models—XCeption, VGG, and ResNet—to assess their effectiveness in addressing this technical challenge.

As we conclude our study, we emphasize the ongoing need for attention and innovation in deepfake detection. XCeption, VGG, and ResNet serve as crucial tools in combating the proliferation of deepfake content, with their performance significantly impacting the security of digital media platforms, journalism, and other sectors. Our findings also demonstrate the potential for real-time application of these detection algorithms, such as within social networks, to mitigate the spread of misinformation. However, given the adaptive nature of deepfake creation, continuous advancements in detection methods and strategies are essential.

By providing a comprehensive evaluation of these three methods, our study contributes to the growing body of knowledge on deepfake detection. We hope that the insights gained will inspire further research, development, and innovation in safeguarding the integrity of digital content. As technology evolves, so must our defenses against potential misuse, aiming for a safer and more secure digital society where the boundary between reality and fiction remains distinct.

Looking ahead, future work aims to implement the algorithm demonstrating the best performance in real-time applications, making these detection mechanisms more accessible to the public and mitigating the current risks posed by deepfakes.

References

- Atas, S., Ilhan, I. and Karakse, M. (2022) 'An Efficient Deepfake Video Detection Approach with Combination of EfficientNet and Xception Models Using Deep Learning', in *2022 26th International Conference on Information Technology (IT)*. *2022 26th International Conference on Information Technology (IT)*, Zabljak, Montenegro: IEEE, pp. 1–4. Available at: https://doi.org/10.1109/IT54280.2022.9743542.
- Chang, X. *et al.* (2020) 'DeepFake Face Image Detection based on Improved VGG Convolutional Neural Network', in *2020 39th Chinese Control Conference (CCC)*. *2020 39th Chinese Control Conference (CCC)*, Shenyang, China: IEEE, pp. 7252–7256. Available at: https://doi.org/10.23919/CCC50068.2020.9189596.
- Dolhansky, B. *et al.* (2019) 'The Deepfake Detection Challenge (DFDC) Preview Dataset'. arXiv. Available at: http://arxiv.org/abs/1910.08854 (Accessed: 14 April 2024).
- Dolhansky, B. *et al.* (2020) 'The DeepFake Detection Challenge (DFDC) Dataset'. arXiv. Available at: http://arxiv.org/abs/2006.07397 (Accessed: 14 April 2024).
- Ghazi, M.M. and Ekenel, H.K. (2016) 'A Comprehensive Analysis of Deep Learning Based Representation for Face Recognition'. arXiv. Available at: http://arxiv.org/abs/1606.02894 (Accessed: 14 April 2024).
- Khatri, N., Borar, V. and Garg, R. (2023) 'A Comparative Study: Deepfake Detection Using Deep-learning', in *2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. *2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India: IEEE, pp. 1–5. Available at: https://doi.org/10.1109/Confluence56041.2023.10048888.
- Mukti, I.Z. and Biswas, D. (2019) 'Transfer Learning Based Plant Diseases Detection Using ResNet50', in *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*. *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, Khulna, Bangladesh: IEEE, pp. 1–6. Available at: https://doi.org/10.1109/EICT48899.2019.9068805.
- Pan, D. *et al.* (2020) 'Deepfake Detection through Deep Learning', in *2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*. *2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*, Leicester, UK: IEEE, pp. 134–143. Available at: https://doi.org/10.1109/BDCAT50828.2020.00001.
- Pashine, S. *et al.* (2021) 'Deep Fake Detection: Survey of Facial Manipulation Detection Solutions'. arXiv. Available at: http://arxiv.org/abs/2106.12605 (Accessed: 14 April 2024).