Evaluating the Effectiveness of Deepfake Video Detection Tools: A Comparative Study

Miroslav Ölvecký [1], Ladislav Huraj [1], Ivan Brlej [1]

[1] *Institute of Computer Technologies and Informatics, University of Ss. Cyril and Methodius in Trnava, Námestie J. Herdu 2, SK-917 01, Trnava, Slovak Republic*

Abstract **– This paper focuses on the evaluation of selected tools for the detection of deepfake videos, which pose a growing threat to the integrity of digital information and the trustworthiness of online media. With the increasing availability of artificial intelligence to create highly realistic manipulated content, the need for robust detection systems is important not only in digital forensics, but also in the broader fields of information security and media verification. This study provides a comparative analysis of five deepfake detection tools, including three open source tools (SBI, LSDA, Lipinc) and two commercial solutions (Bio-ID, Deepware), tested on a dataset of 300 manipulated videos from Celeb-DF (v2). The results indicate a better performance of the commercial tools, with Bio-ID achieving a detection accuracy of 98.00% and Deepware 93.47%, outperforming the open source alternatives.**

The broader implications of this research highlight its potential to strengthen digital trust and combat the spread of disinformation. Reliable detection mechanisms are important for ensuring the authenticity of multimedia content, protecting public figures from attacks on their reputations, and ensuring the credibility of news media.

The findings also highlight the importance of continuous innovation in detection algorithms to respond to the evolving sophistication of deepfake technologies.

This study provides practical insights for developers, researchers, and policymakers to improve detection tools and contribute to a safer digital environment.

Keywords – Deepfake, detection, multimedia file manipulation, disinformation, digital forensics.

1. Introduction

Nowadays, the threat of deepfake multimedia files is increasingly growing, raising concerns about the credibility of media content. Artificial intelligence is often used to create deepfake videos, making it more and more challenging to detect. An example of such a threat is the President of the Slovak Republic, who was recently the victim of a deepfake video. This video, which was posted on the YouTube platform [1], shows how sophisticated this technology can be and how real and convincing the results can be. The deepfake generator in this case used visual information and combined it with an audio voice, but altered the audio information as it would be spoken by the person in the video [2].

In the past, other political figures have been manipulated using similar techniques, not only in the Slovak Republic, but around the world. It is therefore difficult to discern whether the information is real, both, audio and visual content, or if someone is manipulating the information to create misinformation that can lead to social as well as international problems. Therefore, it is important to test different tools for detecting manipulated multimedia files, which are also available to the average Internet user. When testing and evaluating technologies, it is important to consider their applicability and effectiveness in different contexts, for example, as demonstrated by the use of Petri nets for adaptive learning in serious games [3]. Similar to how Petri nets find their application in adaptive learning scenarios, the deepfake video detection technologies use advanced models and algorithms to improve their effectiveness [4].

DOI: 10.18421/TEM141-07
https://doi.org/10.18421/TEM141-07

Corresponding author: Miroslav Ölvecký,
Institute of Computer Technologies and Informatics, University of Ss. Cyril and Methodius in Trnava, Námestie J. Herdu 2, SK-917 01, Trnava, Slovak Republic
Email: miroslav.olvecky@ucm.sk

Received: 17 July 2024.
Revised: 04 December 2024.
Accepted: 24 December 2024.
Published: 27 February 2025.

Deepfake technology is becoming a dangerous tool of war in cyberspace due to its easy availability. Thanks to artificial intelligence and the rapid development of neural networks, there are a number of freely distributable tools for the public to create deepfake content. Even an inexperienced user can create a deepfake of a well-known celebrity, using these tools and a large enough dataset of photos or videos of the person. Technological developments are contributing to more efficient training of neural networks and increased computational power, resulting in more trustworthy results. The development of these tools causes even the smallest details of the images being created to be built more perfectly, thus increasing the quality and authenticity of the artificially created digital content [5].

The manipulation of multimedia content and thus the spread of misinformation places high demands on information security experts in cyberspace, not excluding the digital forensics industry. Therefore, the demand in the field of multimedia file manipulation is growing exponentially, where tools for deepfake detection and detection are trained on datasets [6].

Digital forensics is the scientific field through which experts seek to confirm or disprove the manipulation of digital content in the form of digital footprints, which are required in any investigation where the integrity of digital evidence needs to be preserved. ISO/IEC 27037:2012 [7] is a document that references other ISO/IEC standards in its definition and contains several guidelines and practices in the field of digital forensics. The standard lists the following steps as the basis of the procedure: Identification, acquisition and preservation of digital evidence, extraction, analysis and then production of a final report or forensic report. In terms of deepfake, the extraction phase of digital evidence is important as digital traces can be tampered with in order to destroy the traces, but in this case it is the manipulation of the digital trace that is one of the techniques of antiforensic approaches. Therefore, improper detection of deepfake in the digital trace may prevent the proper recovery of the proven/clarified case [8], [9].

In recent years, there has been a significant increase in the number of manipulated multimedia files, which correlates with a significant increase in cybercrime. The rapid technological advancement of neural networks and deep learning, which contributes to more perfect manipulated files, makes it difficult for forensic experts to detect and forces them to continuously improve their detection techniques and tools. This trend highlights the need to develop specialized efficient tools and train neural network models designed to accurately detect manipulated digital traces, and for this reason, this study focuses on the evaluation of deepfake video detection tools: Bio-ID, Deepware, SBI, LSDA and LIPINC.

It is important to note that such deepfake video detection is available to the average Internet user,

where they can test the accuracy of the multimedia information with the stated accuracy in terms of various generation of manipulated images, audio, etc.

This study contributes to a better understanding of the performance of various available tampered video detection tools and provides a basis for further research and development of effective tools to protect against digital threats. The results of the research also revealed some shortcomings, namely the need to include real videos in the testing to validate the accuracy of the evaluation of detection models and techniques. Also, in terms of investigating real digital traces, the accuracy of the error rate and the evaluation of the integrity of the information being proven is important.

A comparison of different deepfake video detection tools is described, evaluating three open-source tools (SBI, LSDA, and Lipinc) against commercial solutions (Deepware and Bio-ID) on a sample of 300 manipulated videos from the Celeb-DF (v2) dataset. The results of this research contribute to a better understanding of the performance of available deepfake video detection tools and provide a basis for further research and development of effective tools to protect against digital threats.

2. Theoretical Background

Deepfake generation is an innovative technology used for media manipulation. It overcomes the significant shortcomings of traditional forgery methods, by minimizing tampering traces and digital fingerprints that have been used to detect forgeries, it also minimizes inconsistencies in metric or compression artifacts. The functionality of the technology relies on deep neural networks that learn to map segmentation or latent representation to extract input features and then generate new realistic content based on the input data. With the low difference between the boundary of real and fake data, deepfake detection becomes more difficult compared to traditional media manipulation. The basic models for deepfake creation include [10]:

- Auto regression model
- Auto encoder
- Generative Adversarial Network (GAN).

In 2017, a new and more stable generative model architecture was developed in order to increase the overall training stability, this architecture is known as DCGAN, it uses a deep convolutional approach without normalization and batch pooling, which enables better image fusion and arithmetic vector based performance [11].

Later, NVIDIA researchers proposed the ProGAN architecture.

It is a neural network architecture developed to improve the output quality and stability during network training. The architecture includes incremental training on a low-resolution input and then incremental improvement of small details during the training run [12].

The StyleGAN model is based on the previous ProGAN architecture. The developers changed the generator structure with Adaptive Instance Normalization (AdaIN) with the intent of driving generator learning at each convolutional layer. The generator produces a consistent style or position based on the provided vector. A stochastic variation was also created for the position of hair, stubble, skin, etc.

During the course of the research, a problem arose that even when using adaptive normalization, StyleGAN still produced significant artifacts in the synthetic image that created a distracting impression, but the bigger problem was the ease of analyzing these artifacts. Because of this, the developers were forced to redesign the original architecture, therefore, later that year they introduced a new StyleGAN version2 normalization approach that removed the artifacts and distracting impression. ProGAN and StyleGAN are widely used to create synthetic face databases [13], [14].

Figure 1 shows the four most well-known types of deepfake manipulation: Entire Face Synthesis, Reenactment, Facial Attributes Manipulation, Face Swap.

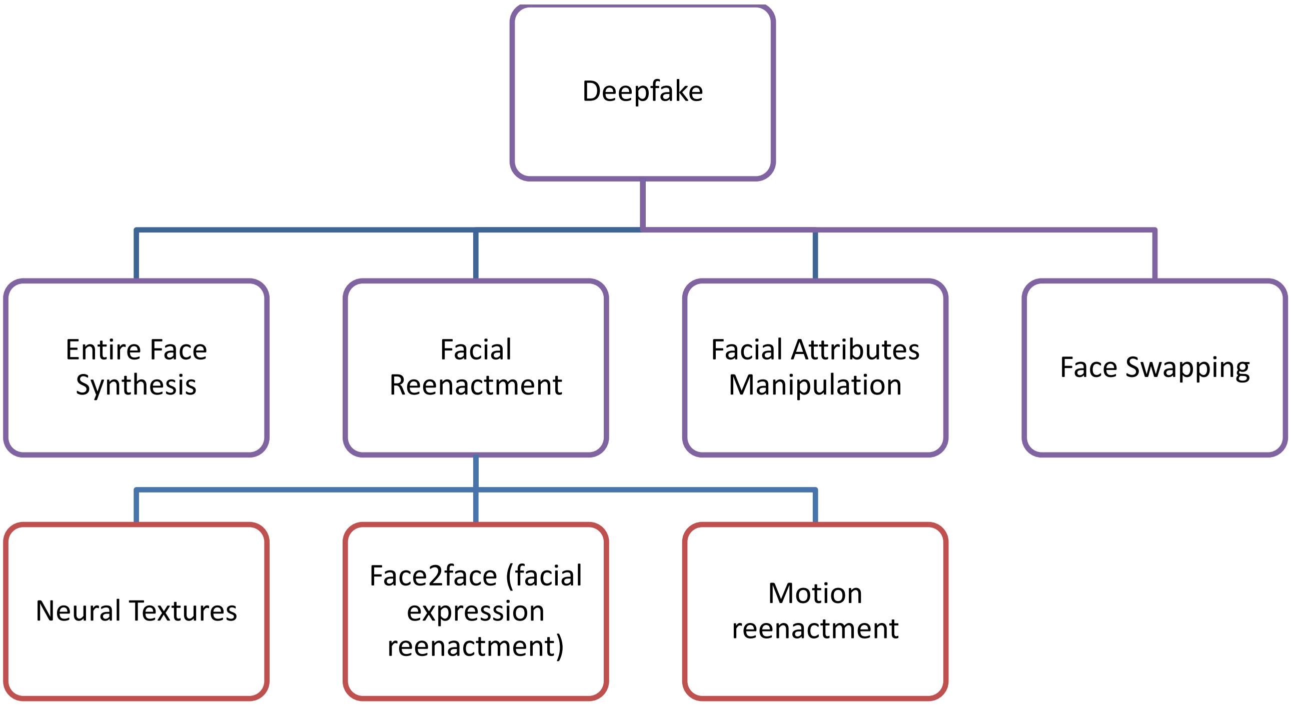


Figure 1. Types of deepfake generation [13]

There are several tools for detecting manipulated videos, which are constantly evolving and improving their accuracy by training models on datasets and evaluating their accuracy. Seow *et al.* [13] define detection tools into two categories (Figure 2): Neural network-based models (CNN, RNN, LSTM) and deep learning techniques.

Heo *et al.* [14] evaluate deepfake detection via a vision tranformer combined with a CNN model. Vision transformers use advanced image processing and learning techniques based on transformer architectures, enabling efficient extraction and analysis of visual features to increase the accuracy and robustness of detection of manipulated visual content [15].

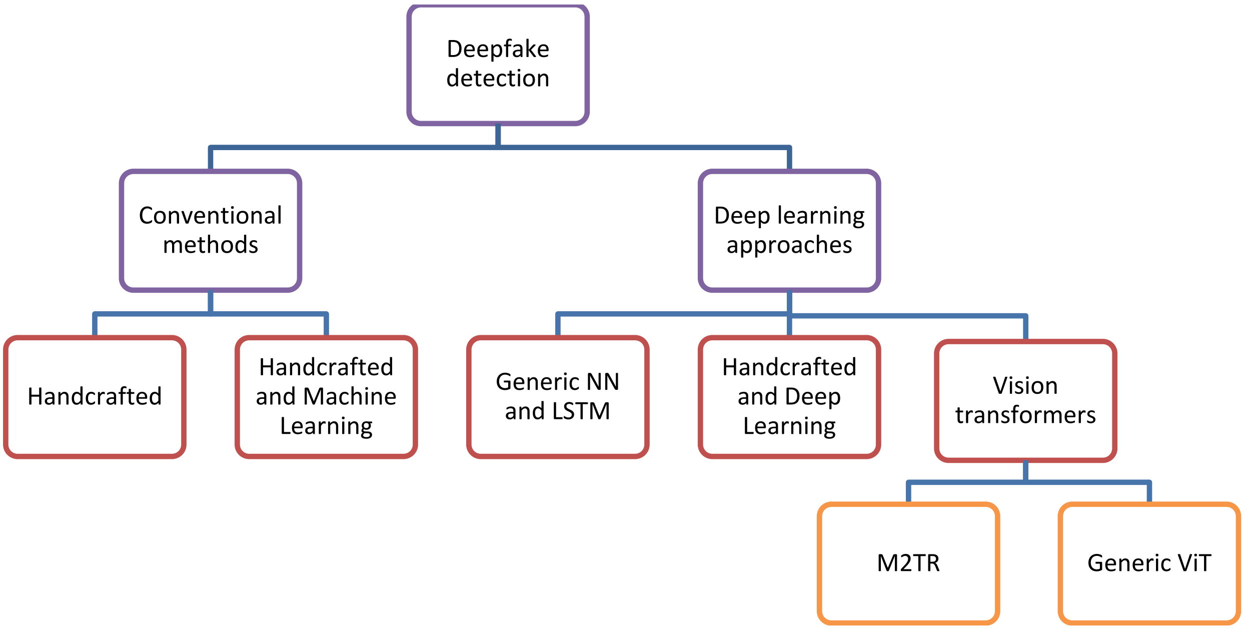


Figure 2. Modified deepfake detection [13]

Since this study was directed at the general Internet user with free access to video manipulation detection tools, the Deepware tool (uses a deep learning model), which is purely focused on video detection, and the Bio-ID tool (uses a deep learning model), which is focused on both video and photo detection, were taken into account in the evaluation. Three other video detection tools were also evaluated, namely SBI (uses CNN), LSDA (uses CNN) and LIPINC (uses a combination of CNN and RNN, possibly also LSTM).

The Deepware tool is designed for deepfake detection via Cloud API, SDK, web interface. Deepware uses the EfficientNet B7 deep learning model, which was trained on the ImageNet dataset. The developers decided to train their own classifier only on Facebook's DFDC dataset, which contains 120,000 videos. The model is a frame-based classifier, which means that it does not consider temporal context. Since video is a temporal medium, the developers consider this a significant drawback that needs to be addressed. The software has been trained on the DFDC dataset, which contains approximately 20,000 real videos, from which 100,000 deepfake videos are subsequently generated using different methods. There are approximately 400 different people in these videos. In most of the videos, a single person is depicted with a length of 10 seconds, while fake versions of the real videos are shown in the metadata. The developers decided to identify unique persons at the beginning of training to prevent identity matching, of which 90% of the persons were used for training and 10% for verification. For testing, they used other scientific datasets focusing on single-person videos from which they extracted faces, using augmentations such as Flip, GaussianNoise, Blur, BrightnessContrast, and many more during training. According to empirical analysis, the developers found that a frequency of 1 FPS represents a trade-off between speed and detection accuracy. Grouping faces allows eliminating noise such as misidentified faces [16].

The Bio-ID tool available as a web interface offers a wide range of features through photo verification, face recognition, photo-biometric template matching and last but not least photo and video tamper detection. The technology uses machine learning algorithms to analyze facial features, gestures and other elements in videos and photos. These algorithms are trained on large datasets of real and synthetic (deepfake) videos. After identifying patterns and anomalies in these datasets, the detection distinguishes between authentic and manipulated content [17].

The LIPINC model can be used as a tool to identify spatio-temporal inconsistencies of the mouth at both local and global levels for deepfake detection. The model analyzes the video for authenticity, evaluating each video using a label $c \in \{0,1\}$, where 0 indicates fake and 1 indicates real. The model consists of two main modules namely, Local and Global Mouth Frame Extractor and Mouth Spatial-Temporal Inconsistency Extractor. In the Local and Global Mouth Frame Extractor module, the face detector is first used to crop and align the face. Subsequently, the face points are utilized to extract the mouth region. The authors propose to extract multiple frames with the mouth open because they contain more inconsistencies related to deepfake. They further analyze the images for local inconsistencies. In addition, they search the rest of the video for other frames with similar mouth position to check for global inconsistencies.

To ensure global comparison and avoid selection of neighboring frames, extractions of similar frames are set to have a minimum time gap of 0.09 seconds between them. The color and structural information from these images are used to extract inconsistent features. The Mouth Spatial-Temporal Inconsistency Extractor module is used to encode color, structural images and learn features that identify spatial and temporal inconsistencies for deepfake detection. The 3D-CNN model to generate spatio-temporal features for color and structural sequences is used. The features are then connected using a cross-attention module that ensures that the branches are correctly merged. Based on the results of the analysis, the LIPINIC authors found that structural features are the key to better deepfake detection performance, so they are given more weight in merging the branch outputs. Finally, the output is used for deepfake prediction using a binary classifier [18], [19].

SBI is a tool with the detection of statistical inconsistency between the modified faces and the background on deepfake. To train robust detectors, the authors of the tool created synthetic fake samples with frequent forgery features that are difficult to identify. Instead of training the models only on existing deepfake videos, they created their own fake samples. Using self-blended images (SBIs), they created synthetic fake samples that combine different features of the source and target images, making detection even more challenging. By doing so, a better and more general face forgery detection on deepfakes is achieved. The source-target generator (STG) works with an input image I, which it copies and creates pseudo source and target images from it. To create differences between them, STG randomly changes their color and frequency values and also scales the source image and shifts it. Mask Generator (MG) provides a grayscale mask to merge the source and target images. As a first step, MG applies a salient point detector to the input image to predict the face region and initializes the mask by computing a convex hull from the predicted salient points of the face. Then, the mask is deformed with a transformation of the significant points as in the BI case (bilinear interpolation). After the first smoothing, pixel values that are less than 1 are reduced to 0. That is, the mask is narrowed if the first Gaussian filter has a larger kernel size than the second one and widened otherwise. Finally, the blending ratio of the source image is changed, allowing the creation of realistic but hard-to-detect false samples, which greatly improves the accuracy and robustness of the detection models [20], [21].

LSDA is a tool based on a heuristic strategy aimed at expanding the falsification space through sample interpolation, which encourages models to learn a more robust decision boundary and to mitigate overfitting to a particular type of falsification.

The authors of the tool propose a latent space augmentation method that allows smooth transitions between different types of falsification without direct dependence on pixel-level artifacts. The use of the Mixup technique in inter-domain augmentation and the ArcFace pre-trained face recognition model contributes to a more robust and comprehensive representation of real faces in the detection model. Experimental results validate the effectiveness of the proposed method, which outperforms current deepfake detectors in generalization between different datasets [22], [23].

3. Methods and Study Design

The video manipulation tools mentioned above have been tested on the available Celeb-DF (v2) dataset [24] which contains 590 original videos collected from YouTube with different age groups, ethnicities, and genders, and 5639 DeepFake videos. The average video length is approximately 13 seconds with 30 FPS (frames per second). The real videos are obtained from publicly available YouTube content, specifically from interviews with 59 celebrities. Among them, 56.8% are male and 43.2% are female. In terms of age, 8.5% of the subjects are aged 60 years and above, 30.5% are aged between 50 and 60 years, 26.6% are aged 40 years, 28.0% are aged 30 years, and 6.4% are younger than 30 years. The ethnic composition of the dataset 5.1% are Asian, 6.8% are African American, and the majority, 88.1%, are Caucasian. Real-life videos vary widely in a variety of factors such as face size, orientation, lighting conditions, and background. As for the DeepFake videos, they are created by exchanging faces among 59 subjects and are delivered in MPEG 4.0 format [25].

For the purpose of this study, 300 deepfake videos were randomly selected from the Celeb-DF (v2) dataset and subsequently tested using Bio-ID, Deepware, SBI, LIPINC and LSDA. For simplicity, the detection threshold was set to 0.6. This value was chosen to provide an optimal balance between sensitivity and detection accuracy across all tools tested. Additionally, a threshold of 0.4 was introduced for categorizing videos as "suspicious". This lower threshold allows for the identification of videos that exhibit some signs of tampering, but not enough to clearly classify them as deepfake. In this manner, a broader spectrum of potentially manipulated videos that necessitate further scrutiny can be captured. Similar approaches have been discussed in the literature, where different thresholds allow finer distinctions between categories to improve detection and analysis of potential threats [26].

The Bio-ID tool provided a binary rating of videos, classifying them as "fake" or "not fake".

On the other hand, the other four tools returned the percentage probability that a video is a deepfake. Based on these percentages and the thresholds set, the videos for each tool were classified into three categories: "Fake Video" (if the percentage exceeded the threshold of 0.6), "Suspicious Video" (if the percentage was between 0.4 and 0.6), and "Real Video" (if the percentage was less than 0.4). Only DeepFake videos from the Celeb-DF (v2) dataset were selected for testing purposes. This approach allowed a clear assessment of the performance of the tested tools, as each video was known to be fake, making it easier to identify where the detection tool had made a mistake. In this way, it was possible to determine exactly when and how often each tool misclassified DeepFake videos as real or merely suspicious, reflecting the accuracy and reliability of the tools.

The second way of evaluating the tools was to consider the exact percentage estimates of the video modification rate reported by each tool. Since the Bio-ID tool does not provide percentage values, the following procedure was established for its results: The output "fake" was assigned a value of 100% and the output "not fake" was assigned a value of 0%. It should be noted that such a simplification may lead to biased results, as it does not account for the finer nuances and degree of uncertainty that may be present when evaluating videos with other tools. However, this approach allowed for a consistent comparison of the performance of all tools tested within the same evaluation framework.

Subsequently, various types of comparisons, including descriptive statistics, frequency analysis, data visualization, and Related-Samples Friedman's Two-Way Analysis of Variance by Ranks (Friedman's test), were conducted for each method to reveal differences and similarities between tools. Descriptive statistics provided a basic overview of the data, including means, medians, and variances, allowing for an initial view of the distribution and characteristics of the data being assessed. Frequency analysis was used to identify frequent and recurring patterns in the data, contributing to a better understanding of the distribution of individual instrument results. Visualization of the data, through various graphs and charts, allows for a visual representation of the statistical patterns found and the variation between instruments. Friedman's test as a statistical test suitable for comparing multiple related samples was used to identify significant differences between detection tools.

Statistical tests were performed using IBM SPSS software version 28.

4. Results

The goal of this study was to analyze the success of various video modification detection tools. Two different methodologies were used for this purpose. The first methodology consisted of thresholding the tools' responses into categories where videos were classified as "Fake Video", "Suspicious Video" or "Real Video" based on set thresholds. The second methodology took into account percentage ranking, where each video was assigned an exact percentage value for the probability of being a deepfake.

Based on these methodologies, the results of the tool evaluations were created to provide insight into the performance of the tools. The threshold ranking allows for a simple and straightforward comparison between tools, while the percentage ranking provides deeper insight into each tool's ability to identify different levels of modification.

4.1. Results of Software Comparison Based on Threshold Ranking

Based on the outcomes obtained from each tool, the individual outputs were coded on a four-point scale: A value of 0 corresponded to an error wherein the tool was unable to process or analyze the video input, thereby failing to produce any meaningful evaluation result; A value of 1 corresponded to the tool incorrectly indicating in response to the video that video was real (not fake); A value of 2 was assigned to the response that the video was suspicious; and a value of 3 was assigned to the correct response that the video was deepfake. This scale was selected because the primary objective of the evaluated tools is to accurately identify manipulated videos, which is the most important and definitive outcome for assessing the tools' performance.

The average value graph shows the average score that each tool achieved in detecting video modifications, Figure 3. Higher average scores indicate a tool's better ability to identify deepfake videos. From the graph, it can be seen that the Bio-ID and Deepware tools achieved the highest mean values, 2.96 and 2.97 respectively, indicating their high effectiveness in detecting deepfake videos. The SBI tool achieved an average value of 2.88, which equally ranks it among the best tools tested. The LIPINC and LSDA tools achieved lower mean values, 2.70 and 2.64 respectively, indicating that their ability to identify deepfake videos is less conclusive compared to the other tools.

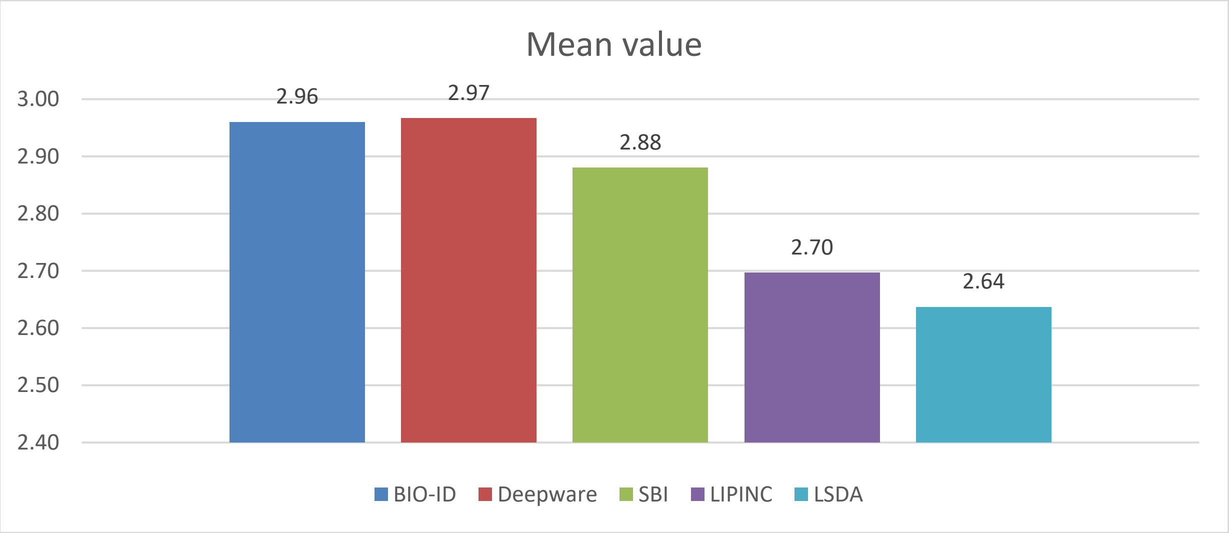


Figure 3. Graph of the average value for the tested instruments

Table 1. Descriptive statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Bio-ID	300	1	3	2.96	.280
Deepware	300	1	3	2.97	.243
SBI	300	1	3	2.88	.382
LIPINC	300	0	3	2.70	.707
LSDA	300	0	3	2.64	.673
Valid N (listwise)	300				

The descriptive statistics presented in Table 1 record more detailed values of each deepfake detection tool.

Based on descriptive statistics, it can be concluded that:

- Deepware has the highest mean (2.97) and the lowest standard deviation (0.243), indicating that it is the best rated and the ratings are very consistent.

- Bio-ID also has a high mean value (2.96) and a relatively low standard deviation (0.280), indicating that it is very well rated.

- The SBI has a mean value of 2.88, which is still relatively high, but with more variance (0.382) in the ratings.

- LIPINC and LSDA have lower means (2.70 and 2.64) and higher standard deviations (0.707 and 0.673), indicating greater variance in ratings and less consistent ratings.

A graph of the frequency of measured values provides an overview of the number of category outputs for each instrument (Figure 4). This graph visually compares how many times each tool was able to correctly identify videos as "Fake", "Suspicious" or "Real", and how many times an error occurred. The graph shows that the Bio-ID and Deepware tools had the highest number of correctly identified "Fake" videos, with 294 correctly identified videos, indicating their high accuracy. The SBI tool achieved 270 correct identifications, ranking it as an effective tool, although with a slightly lower number of correct identifications compared to Bio-ID and Deepware. On the other hand, the LIPINC and LSDA tools showed a higher number of errors (Error) and a lower number of correctly identified fake videos, with values of 251 and 223 for fake videos, respectively, indicating lower accuracy and reliability of these tools in detecting deepfake videos, Table 2.

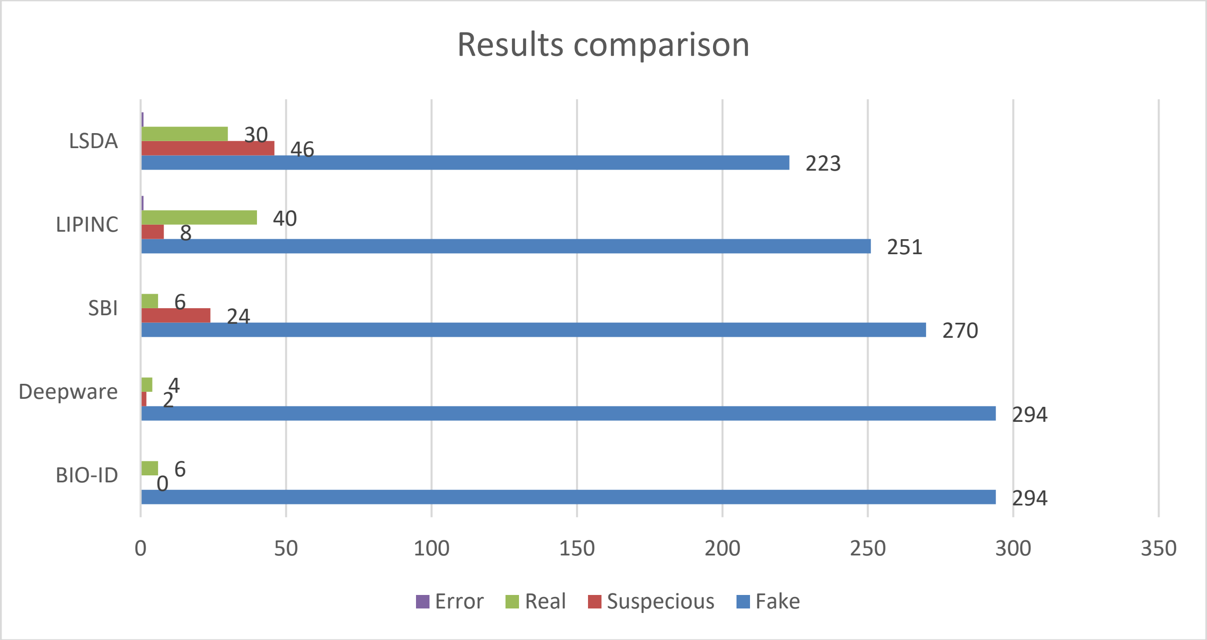


Figure 4. Graph of frequency of measured values

Table 2. Frequency of measured values

Tool	Fake (Count, %)	Real (Count, %)	Suspicious (Count, %)	Error (Count, %)
Bio-ID	294 (98.00%)	6 (2.00%)	0 (0.00%)	0 (0.00%)
Deepware	294 (98.00%)	4 (1.33%)	2 (0.67%)	0 (0.00%)
SBI	270 (90.00%)	6 (2.00%)	24 (8.00%)	0 (0.00%)
LIPINC	251 (83.67%)	40 (13.33%)	8 (2.67%)	1 (0.33%)
LSDA	223 (74.33%)	30 (10.00%)	46 (15.33%)	1 (0.33%)

Related-Samples Friedman's Two-Way Analysis of Variance by Ranks, known as Friedmann's test, was used to compare the performance of different deepfake video detection tools. The aim was to test the null hypothesis that "the distributions of the outputs of the Bio-ID, Deepware, SBI, LIPINC and LSDA tools are the same".

The results of the test showed that the null hypothesis can be rejected with a significance level of 0.050, which means that there are statistically significant differences between the outputs of each tool, Table 3.

Table 3. Friedmann test results

Total N	300
Test Statistic	119.982
Degree Of Freedom	4
Asymptotic Sig.(2-sided test)	.000

Pairwise comparisons were performed to identify specific differences between pairs of instruments. Each row in the table tests the null hypothesis that the distributions of Sample 1 and Sample 2 are the same.

Significance (Sig.) and adjusted significance (Adj. Sig.) by Bonferroni correction are reported for each comparison. Adjusted significance takes into account multiple comparisons.

Table 4. Friedmann test with pairwise comparisons

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
LSDA-LIPINC	.222	.129	1.717	.086	.860
LSDA-SBI	.395	.129	3.060	.002	.022
LSDA-Bio-ID	.587	.129	4.544	<.001	.000
LSDA-Deepware	.588	.129	4.557	<.001	.000
LIPINC-SBI	.173	.129	1.343	.179	1.000
LIPINC-Bio-ID	.365	.129	2.827	.005	.047
LIPINC-Deepware	.367	.129	2.840	.005	.045
SBI-Bio-ID	.192	.129	1.485	.138	1.000
SBI-Deepware	.193	.129	1.498	.134	1.000
Bio-ID-Deepware	-.002	.129	-.013	.990	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is .050.

Table 5. Results of pairwise comparisons from the Friedmann test for individual instruments

Tool	Bio-ID	Deepware	SBI	LIPINC	LSDA
Bio-ID	-	Not Significant	Not Significant	Significant	Significant
Deepware	Not Significant	-	Not Significant	Significant	Significant
SBI	Not Significant	Not Significant	-	Not Significant	Significant
LIPINC	Significant	Significant	Not Significant	-	Not Significant
LSDA	Significant	Significant	Significant	Not Significant	-

Significant means that there is a statistically significant difference between a pair of software.
Not Significant means that there is no statistically significant difference between the pair of software.

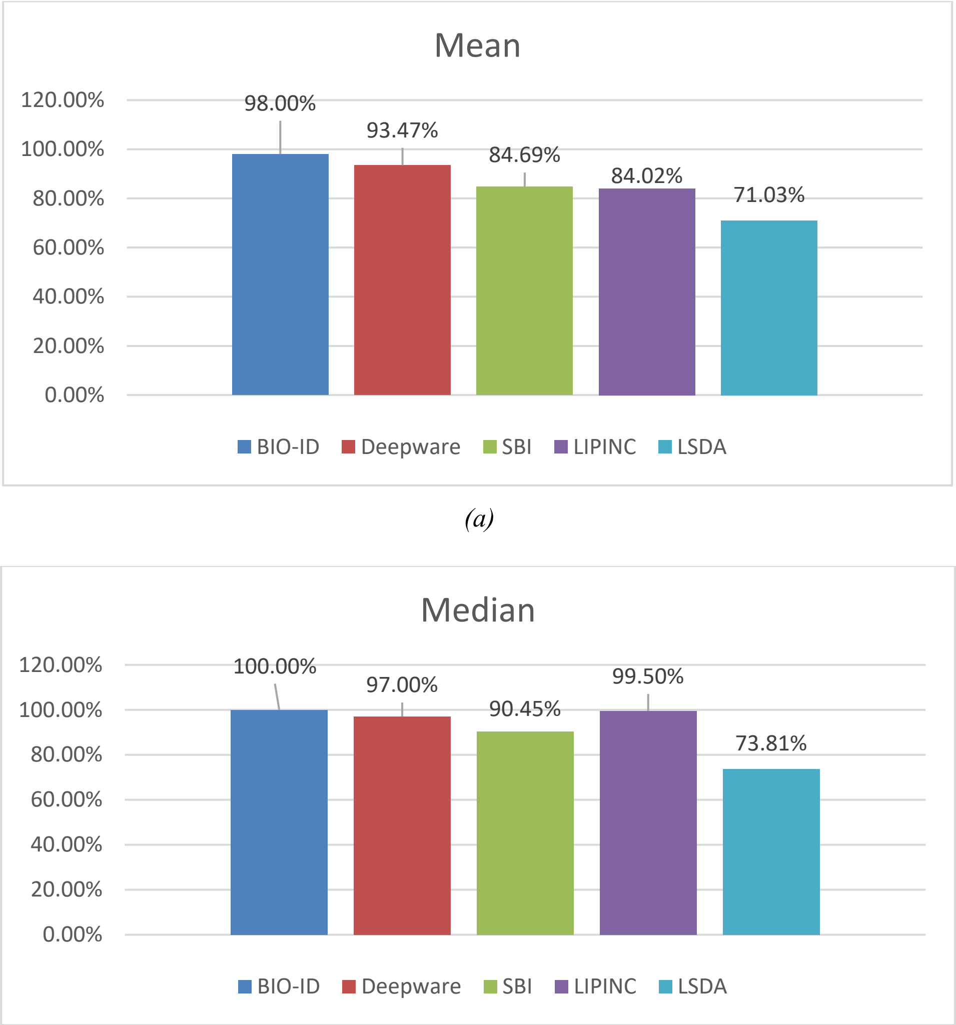
The results of the Friedmann test with pairwise comparisons indicate whether there are significant differences between the columns Table 5.

Pairwise comparisons show that LSDA is statistically significantly different from Bio-ID, Deepware, and SBI, while Bio-ID and Deepware, Bio-ID and SBI, Deepware and SBI, LIPINC and SBI, LIPINC and LSDA, and Bio-ID and Deepware are not statistically significantly different.

4.2. Results of Software Comparison Based on Percentage Results

The second methodology took into account the percentage ranking, whereby each video was assigned an exact percentage value for the probability of being a deepfake. Using the assumption that the Bio-ID tool gives a 100% value if it correctly identifies a video as "Fake" and a 0% value if it incorrectly identifies a video was acceptable in this case as a simple and straightforward method.

The graphs in Figure 5 show the mean and median percentages of deepfake video detection success for each of the tools evaluated, providing a visual overview of the performance of each tool across the test sample of 300 videos.



(a)

(b)

Figure 5. (a) Graph of mean for percentage rank, (b) Graph of median for percentage rank

Also in the second methodology, descriptive statistics were used to analyse the performance of individual instruments in detail.

Table 6 summarizes the basic statistical parameters including minimum, maximum, mean, median, and standard deviation for each instrument, providing deeper insight into their performance.

Table 6. Descriptive statistics

Tool	N	Minimum	Maximum	Average	Median	Std. Deviation
Bio-ID	300	0.00%	100.00%	98.00%	100.00%	14.02%
Deepware	300	29.00%	98.00%	93.47%	97.00%	8.90%
SBI	300	30.62%	99.97%	84.69%	90.45%	16.11%
LIPINC	300	0.00%	100.00%	84.02%	99.50%	30.19%
LSDA	300	0.00%	99.24%	71.03%	73.81%	17.99%

Figure 6 shows the variance of the scores for each tool, where it can be seen that Bio-ID and Deepware are the most reliable tools with the highest consistency, while SBI, LSDA and LIPINC show a higher level of variability and less reliable results.

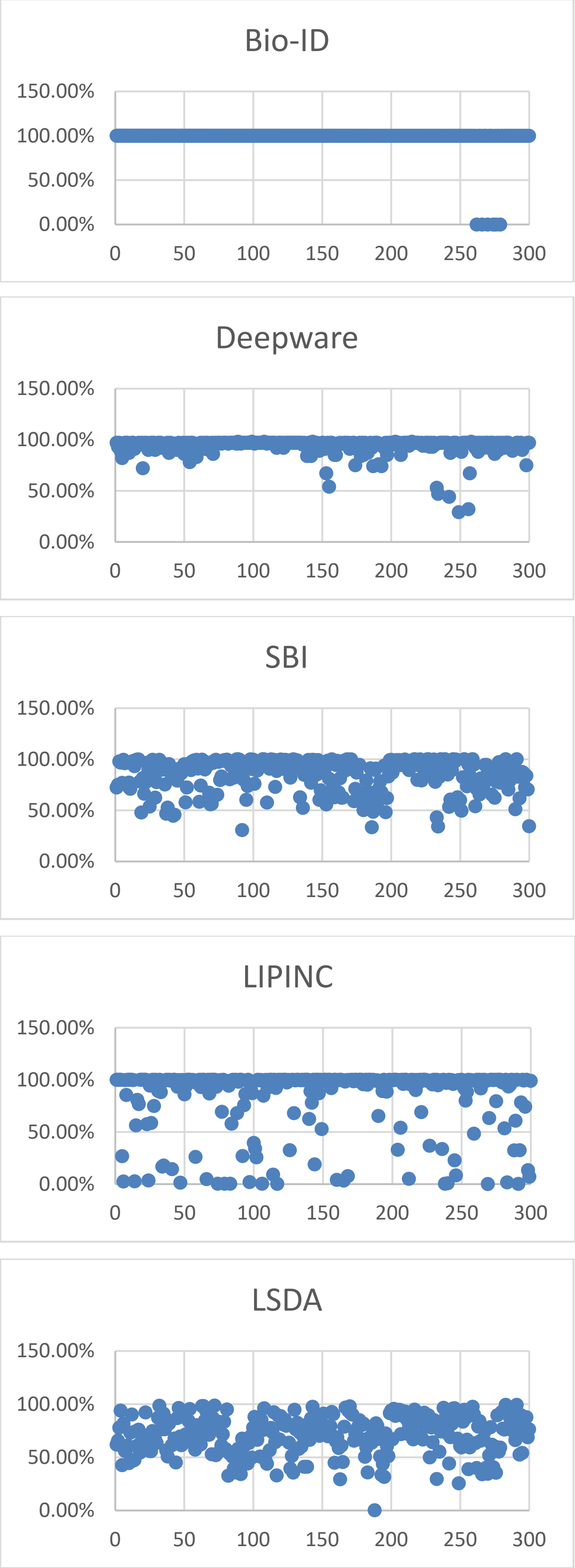


Figure 6. Variability and consistency of tool results

From the measured values, it can be seen that the Bio-ID tool is the most successful deepfake video detection tool among the tested tools with the highest average success rate of 98.00% and median success rate of 100.00%. This means that Bio-ID is very consistent and reliable in detection with high success rate. It is followed by Deepware tool as the second most successful software with an average success rate of 93.47% and a median success rate of 97.00%. It has a low standard deviation, indicating high consistency of results. SBI has an average success rate of 84.69% and a median success rate of 90.45%. This tool is less consistent compared to Bio-ID and Deepware, indicating higher variability in success rates. LIPINC has an average success rate of 84.02% and a very high median success rate of 99.50%. The high standard deviation (30.19%) indicates that the success rate of this software is highly variable and less reliable. LSDA has the lowest mean success rate of 71.03% and a median success rate of 73.81%. This suggests that LSDA is the least successful software of those analyzed with significant variability in success rates.

From the analysis, it is clear that Bio-ID is the most successful tool among the video modification detection tools tested, followed by Deepware. SBI and LIPINC have similar average success rate, but LIPINC shows more variability. LSDA is the least successful software in this comparison.

However, it should be repeated that the Bio-ID software does not give percentages in the calculation when detecting a modification and therefore they were designed as 100% for recognizing a modification and 0% for not recognizing a modification. This may affect the overall statistics and analysis of success rates as it does not account for more subtle differences in detection and may bias the results in favour of the Bio-ID software compared to other software that provides more accurate percentage success rates.

5. Discussion

This study analyzed the performance of five deepfake video detection tools: Bio-ID, Deepware, SBI, LIPINC and LSDA. The results provided initial insight into the ability of these tools to correctly identify deepfake videos using two different methodologies. The Bio-ID and Deepware tools proved to be the most effective for both methodologies.

For thresholding, each video was assigned a category of "Fake Video", "Suspicious Video" or "Real Video" based on the thresholds set. This approach allowed for a simple and straightforward comparison between tools.

The results showed that Bio-ID and Deepware achieved the highest numbers of correctly identified fake videos (Fake), with values of 294 out of 300, corresponding to a success rate of 98.00%. SBI, with a value of 270 correctly identified videos (90.00%), also proved to be an effective tool, although less consistent than Bio-ID and Deepware.

On the other hand, LIPINC and LSDA achieved lower success rates, with values of 83.67% and 74.33%, respectively, indicating lower reliability of these tools.

The second methodology took into account an accurate percentage assessment of the probability that a video is a deepfake. This approach allowed a more detailed assessment of the tools' ability to detect different levels of video modification. The results showed that Bio-ID and Deepware achieved the highest average success rates, with values of 98.00% and 93.47%, respectively. In contrast, the SBI, LIPINC and LSDA tools showed a greater variability in success rates, with values of 84.69%, 84.02% and 71.03%, respectively.

Bio-ID and Deepware are commercial tools. These tools demonstrated the ability to correctly identify fake videos with average success rates of 98.00% and 93.47%, respectively. The advantage of commercial tools is that they are often backed by extensive development, testing, and technical support. Their use is usually simple and intuitive, allowing even less experienced users to effectively detect deepfake videos. However, the disadvantage can be higher costs and limited access for individuals or smaller organizations. The Deepware tool showed slightly better results of 93.47% in the study than the 85.60% results reported in [27], although this was not the same Celeb-DF (v2) dataset, but a similar Celeb-DF Fake dataset.

The SBI, LIPINC and LSDA tools are non-commercial tools that were developed in an academic environment. These tools showed more variability in the success rate of detecting deepfake videos. The average success rates for SBI, LIPINC, and LSDA were 84.69%, 84.02%, and 71.03%, respectively. The variability in results was higher, indicating less consistent performance compared to commercial tools. The advantage of non-commercial tools is their accessibility to the general public and the ability to customize and refine as needed. Disadvantages may be lower reliability and lack of technical support, which may limit their use for general users.

This is consistent with Almars [28], who argues that although machine learning techniques have demonstrated remarkable performance in detecting deepfake videos, the quality of deepfake videos is constantly increasing; therefore, there is a need to continuously improve current detection methods in order to successfully identify fake videos in the future.

The findings of this study are especially beneficial for ordinary users who want to verify the authenticity of videos.

The study provides information on the performance of different deepfake video detection tools and shows which tools to use or with what level of confidence they can trust the results of these tools.

Commercial tools such as Bio-ID and Deepware are ideal for ordinary users looking for reliable and intuitive video authentication solutions. Their high accuracy and ease of use make them a suitable choice for individuals and smaller organizations that need to detect deepfake videos quickly and efficiently.

Non-commercial tools such as SBI, LIPINC and LSDA provide an alternative for users looking for cheaper or free solutions. While these tools may have a higher variability in success rates, they can still be useful for authenticating videos, especially when used in combination with other methods or tools.

One significant limitation of this research is the way in which the outputs of the Bio-ID tool are evaluated, which does not report results in percentage form. For the purposes of this study, it was assumed that the Bio-ID tool provides binary ratings: 100% for correctly identified deepfake videos and 0% for misidentified videos. This approach may bias the results because it does not account for finer differences in detection and may lead to an overestimation of the tool's success rate. Another limitation is the use of only one dataset, namely Celeb-DF (v2), which may not adequately represent the diversity and complexity of all types of deepfake videos. Another limitation is the size and composition of the test sample, which consisted of 300 videos, which may be insufficient to generalize the results to a broader population of deepfake videos. In future research, it would be advisable to include a larger number of videos, a larger number of datasets, and different types of deepfake techniques to improve the robustness and generalizability of the results.

6. Conclusion

This study analyzed the performance of five deepfake video detection tools: Bio-ID, Deepware, SBI, LIPINC and LSDA on a sample of fake videos only. The commercial tools Bio-ID and Deepware demonstrated the highest success rate in identifying fake videos with mean values of 98.00% and 93.47%, respectively. The non-commercial and academic tools SBI, LIPINC and LSDA showed more variability in success rates, with mean values of 84.69%, 84.02% and 71.03% respectively, indicating less consistent results. These differences highlight the need for continuous development and improvement of detection algorithms.

The results of this study are beneficial for ordinary users who need to authenticate videos. The limitations of this study are the evaluation of the outputs of the Bio-ID tool, which does not provide results in percentage form, and the use of only one dataset, namely Celeb-DF (v2), which may not adequately represent the diversity of deepfake videos. Future research should include a larger number of videos and different types of deepfake techniques to improve the robustness and generalizability of the results.

Acknowledgements

This research was funded by the grant KEGA 011UCM-4/2024 Teaching and training IoT security using modern technologies.

References:

[1]. YouTube. (2023). *Deepfake s P. Pellegrinim - podvodné investovanie.* YouTube. Retrieved from: https://www.youtube.com/watch?v=N2IFHycwh80 [accessed: 05 June 2024]

[2]. TA3. (2023). *Sociálnymi sieťami kolujú falošné videá s Pellegrinim. Prezidentská kancelária hovorí o podvode, chce zakročiť.* TA3. Retrieved from: https://www.ta3.com/clanok/947165/socialnymi-sietami-koluju-falosne-videa-s-pellegrinim-prezidentska-kancelaria-hovori-o-podvode-zvazuje-pravne-zakrocit [accessed: 06 June 2024]

[3]. Hosťovecký, M., Korečko, Š., & Sobota, B. Petri nets for Adaptive learning scenarios in Serious games. *Journal of Applied Mathematics, Statistics and Informatics*, *20*(1), 67-84.

[4]. Valentová, M., & Brečka, P. (2023). Assessment of Digital Games in Technology Education. *International Journal of Engineering Pedagogy*, *13*(2).

[5]. Tolosana, R., et al. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, *64*, 131-148.

[6]. Lin, H., et al. (2023). DeepFake detection with multi-scale convolution and vision transformer. *Digital Signal Processing*, *134*, 103895.

[7]. ISO/IEC. (2012). *ISO/IEC 27037:2012 – Information technology – Security techniques – Guidelines for identification, collection, acquisition, and preservation of digital evidence*. ISO27001security. Retrieved from: https://www.iso27001security.com/html/27037.html [accessed: 08 June 2024].

[8]. González Arias, R., et al. (2024). Systematic Review: Anti-Forensic Computer Techniques. *Applied Sciences*, *14*(12), 5302.

[9]. Zakaria, S. N. A. B. S., Chao, K. F., & Zainol, Z. (2023). Exploring Data Wiping Practices in the Royal Malaysian Air Force (RMAF) HQ. *International Visual Informatics Conference*, 328-338.

[10]. Patel, Y., et al. (2023). Deepfake generation and detection: Case study and challenges. *IEEE Access*. Doi: 10.1109/ACCESS.2023.3342107

[11]. Yu, Y., et al. (2017). Unsupervised representation learning with deep convolutional neural network for remote sensing images. *Lecture Notes in Computer Science.* Springer.

[12]. Ruiz-Casado, J. L., Molina-Cabello, M. A., & Luque-Baena, R. M. (2024). Enhancing Histopathological Image Classification Performance through Synthetic Data Generation with Generative Adversarial Networks. *Sensors*, *24*(12), 3777.

[13]. Seow, J. W., et al. 2022). A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing*, *513*, 351-371.

[14]. Heo, Y. J., et al. (2021). Deepfake detection scheme based on vision transformer and distillation. *arXiv preprint arXiv:2104.01353*. Retrieved from: https://arxiv.org/abs/2104.01353 [accessed: 05 July 2024].

[15]. Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[16]. Deepware. (2021). *What you see is no longer the truth.* Cmd.deepware. Retrieved from: https://cmd.deepware.ai/downloads/Whitepaper_v1.1.pdf [accessed: 06 July 2024].

[17]. Bio-ID. (n.d.). *Deepfake Detection Software.* Bioid. Retrieved from: https://www.bioid.com/deepfake-detection/ [accessed: 06 July 2024]

[18]. Datta, S. K., Jia, S., & Lyu, S. (2024). Exposing Lip-syncing Deepfakes from Mouth Inconsistencies. *arXiv preprint arXiv:2401.10113*. Retrieved from: https://arxiv.org/html/2401.10113v1 [accessed:07 July 2024].

[19]. Datta, S. K. (n.d.). *Lipinc detection deepfake tool.* Github. Retrieved from: https://github.com/skrantidatta/LIPINC/blob/main/model.py [accessed: 08 July 2024]

[20]. Shiohara, K., & Yamasaki, T. (2022). Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18720-18729.

[21]. SCLBD. (n.d.). *SBI detection deepfake tool.* Github. Retrieved from: https://github.com/SCLBD/DeepfakeBench/blob/main/training/detectors/sbi_detector.py [accessed: 10 July 2024].

[22]. Yan, Z., et al. (2023). Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. *arXiv*. Retrieved from: https://arxiv.org/pdf/2311.11278 [accessed: 10 July 2024].

[23]. SCLBD. (n.d.). *LSDA detection deepfake tool*. Github. Retrieved from: https://github.com/SCLBD/DeepfakeBench/blob/main/training/detectors/lsda_detector.py [accessed: 15 July 2024].

[24]. Varghese, R. S. & Reji, S. (n.d.). *Celeb-DF (v2) dataset*. Kaggle. Retrieved from: https://www.kaggle.com/datasets/reubensuju/celeb-df-v2 [accessed: 15 July 2024].

[25]. Li, Y., et al. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3207-3216.

[26]. Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE access*, *10*, 25494-25513.

[27]. Mukta, M. S. H., Ahmad, J., Raiaan, M. A. K., Islam, S., Azam, S., Ali, M. E., & Jonkman, M. (2023). An investigation of the effectiveness of deepfake models and tools. *Journal of Sensor and Actuator Networks*, *12*(4), 61.

[28]. Almars, A. M. (2021). Deepfakes detection techniques using deep learning: a survey. *Journal of Computer and Communications*, *9*(5), 20-35.