DOI: http://dx.doi.org/10.26483/ijarcs.v17i3.7514

Volume 17, No. 3, May-June 2026

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

EVALUATING PUBLICLY ACCESSIBLE DEEPFAKE DETECTION SYSTEMS: ESTABLISHING PRACTICAL RELIABILITY BENCHMARKS FOR REAL-WORLD IMAGE-BASED DEEPFAKES

Christina Daffi J
St. Joseph's Boys' High School
Bengaluru, Karnataka, India

Abstract: Deepfake technology has rapidly advanced in recent years, raising significant concerns about misinformation, identity manipulation, and the misuse of synthetic media. Although many deepfake detection systems reported in academic literature demonstrate strong benchmark accuracy, relatively little research focuses on the practical reliability and accessibility of publicly usable detection tools for ordinary users. This study addresses that gap through a practical evaluation of publicly accessible deepfake detection systems under benchmark and real-world conditions.

A preliminary accessibility survey of publicly referenced deepfake detection platforms was first conducted. Several systems were found to be restricted, unstable, enterprise-focused, video-only, or unavailable for consistent public experimentation. As a result, Hive Moderation was selected as the primary detector for evaluation. The methodology included evaluating benchmark image samples from FaceForensics++ and Celeb-DF, comparing detector performance with human participants, and testing real-world deepfake images collected from online sources.

The results demonstrated a noticeable performance decline on more realistic, real-world deepfakes. Human participants and the evaluated detector exhibited varying performance across benchmark datasets, while both showed reduced accuracy on Celeb-DF and real-world samples. In addition, several incorrect classifications occurred with extremely high confidence scores, suggesting inconsistency in prediction certainty and practical reliability.

Keywords: Deepfake Detection, Publicly Accessible AI detectors , Digital Forensics , FaceForensics++ , Celeb-DF , Human vs AI detection.

INTRODUCTION

Deepfake technology has rapidly advanced in recent years due to improvements in artificial intelligence, generative adversarial networks (GANs), and diffusion-based image generation systems. These technologies are capable of producing highly realistic synthetic media [1], including manipulated images, videos, and audio recordings that are often difficult for humans to distinguish from authentic content. While deepfake technology has legitimate applications in entertainment, education, and accessibility, it has also introduced significant concerns related to misinformation, identity manipulation, fraud, political disinformation, and digital impersonation [1].

The increasing accessibility of generative AI tools has contributed to the widespread circulation of synthetic media across online platforms and social networks. As a result, deepfake detection has become an important research area within digital forensics and artificial intelligence.

Many deepfake detection systems have been proposed in academic literature, including CNN-based, frequency-based, transformer-based, and hybrid approaches [2, 3]. Several studies report high benchmark accuracies under controlled experimental conditions [4]. However, recent research has shown that detector performance may decline when evaluated under real-world conditions involving compression, resizing, reposting, and other image transformations [5, 6, 7, 8].

Despite the growing body of deepfake detection research, relatively little attention has been given to the practical accessibility and reliability of publicly available detection tools for ordinary users [2]. Many publicly referenced systems are restricted, paid, unavailable, or difficult to access, while existing evaluations often assume that usable public detectors are readily available. In addition, few studies compare publicly accessible detector performance with human participants using both benchmark datasets and real-world online deepfake content [2, 5].

Understanding the practical reliability of publicly accessible deepfake detection tools is increasingly important as synthetic media becomes more widespread online. If publicly available detectors are inaccurate, inaccessible, or unreliable under realistic conditions, ordinary users may have limited ability to verify suspicious content. Evaluating these systems, therefore, provides insight into both the effectiveness and the accessibility of current public deepfake defense tools.

Therefore, this study evaluates the reliability of a publicly accessible deepfake detection system using benchmark datasets, real-world deepfake images, and human participant comparisons. Specifically, the study investigates: (1) how reliably publicly accessible detectors perform on benchmark datasets, (2) how detector performance compares with human judgment, (3) whether performance declines on real-world deepfake images, and (4) the practical accessibility challenges associated with publicly available deepfake detection tools.

This study focuses exclusively on image-based deepfakes and does not evaluate video-based detection systems. In addition, the experimental evaluation primarily examines a single publicly accessible detector due to the limited availability of other consistently usable public platforms. To address the research objectives, an accessibility survey of publicly referenced deepfake detection tools was first conducted. Following detector selection, benchmark evaluations were performed using FaceForensics++ and Celeb-DF image datasets, detector performance was compared with human participant classifications, and additional testing was conducted using real-world deepfake images collected from online sources.

METHODS

Research Design

This study employed an observational comparative evaluation design examining the performance of a publicly accessible deepfake detector across benchmark datasets, real-world samples, and human participant assessments.

Detector Selection and Accessibility Survey

A system is considered publicly accessible in this study if it satisfies all of the following conditions:

- Provides web-based access without requiring local installation or setup
- Does not require programming, coding, or model deployment by the user
- Allows immediate inference or prediction after uploading an input
- Does not require computational infrastructure (e.g., GPU setup, training environment)
- Is usable through a standard browser interface intended for general users
- Is either free to use or provides clearly accessible usage without technical or institutional barriers that restrict ordinary user access

This definition ensures that the evaluation focuses on tools that are realistically usable by non-technical users attempting to verify synthetic media in everyday contexts.

This study initially aimed to evaluate multiple publicly accessible deepfake detection systems. Multiple tools discussed in existing literature and online resources were evaluated for practical accessibility, usability, and functionality. However, several systems were found to have restricted access, paid limitations, unavailable public demos, inconsistent outputs, upload restrictions, or non-functional interfaces during the evaluation period [2].

Table 1. Public Accessibility of Deepfake Detection Tools

Tool	Accessibility / Limitation
Hive	Publicly accessible, free, functional, and used in this study
DecopyAI	Extreme high-confidence false positives; inconsistent outputs
TruthScan	Limited free credits; AI-image detector rather than deepfake-specific
FaceOnLive	Public demo unavailable during testing
BitMind	High-confidence false negatives; general AI detector
InVID	Primarily video-based forensic analysis tool
FotoForensics	Image forensic analysis tool; not deepfake-specific
Forensically	Manual forensic inspection tool rather than automated detector
Deepware Scanner	Slow processing; video-focused
Sensity AI	Enterprise-focused; not publicly accessible
CloudSEK	Primarily video-based analysis
Reality Defender	Restricted free API limits
Intel FakeCatcher	Video-only and not publicly accessible
Microsoft Video Authenticator	Restricted third-party access only
Pindrop Pulse	Audio/video-focused enterprise system
Others	Additional tools showed similar accessibility, functionality, or usability limitations

As a result, this study focuses primarily on the evaluation of Hive Moderation's publicly accessible AI-generated image detector.

Many publicly available systems identified during the accessibility survey focused primarily on video-based analysis. Video deepfake detection can utilize temporal information such as facial motion patterns, blinking behavior, lip-synchronization consistency, and frame-to-frame artifacts that are unavailable in single images [1, 6]. Because this study focused on image-based deepfake verification, video-only systems were excluded from evaluation.

It is also important to distinguish between general AI-generated content detectors and deepfake-specific detection systems. AI content detectors are designed to identify a broad range of synthetic content, including text-to-image outputs and fully generated artwork, whereas deepfake detectors focus specifically on manipulated or synthetically altered human faces and identities [2]. Therefore, this study

prioritized tools that explicitly claimed deepfake detection capabilities.

Following the accessibility survey, Hive Moderation was selected as the primary detector due to its consistent functionality, public accessibility, and ability to provide probabilistic confidence scores. Other publicly referenced systems were excluded because of restricted access, unavailable demonstrations, unstable outputs, limited free usage, or a lack of image-based deepfake detection support.

Datasets

This study utilized two publicly available benchmark datasets commonly referenced in deepfake detection research: FaceForensics++ and Celeb-DF [1, 6]. These datasets were selected due to their widespread usage in prior literature and their inclusion of both authentic and manipulated facial imagery.

For the benchmark evaluation, the study included both real and deepfake image samples extracted from the selected datasets. In addition, a smaller set of real-world deepfake images collected from publicly circulated online sources was included to evaluate practical detector performance outside controlled benchmark conditions [2].

The experimental evaluation, therefore, consisted of three categories of image data:

- Benchmark real images
- Benchmark deepfake images
- Real-world online deepfake images

Figure 1. Examples of real and deepfake images from FaceForensics++, Celeb-DF, and real-world deepfake samples used in this study.

Procedure & Participants

To compare the performance of automated detectors with human judgment, a human evaluation study was conducted using the same benchmark image sets. Participants were asked to classify each image as either REAL or FAKE based solely on visual inspection [5, 8].

The human evaluation involved 25 adult participants who voluntarily completed an anonymous image-classification survey, conducted using a blinded setup in which participants were not informed of the correct labels during testing. Responses were collected and compared against the ground-truth dataset labels to calculate participant accuracy.

Hive Moderation's publicly accessible AI-generated image detector was used as the primary automated evaluation system in this study. Each image was individually uploaded to the platform, and the returned prediction confidence scores were recorded.

Hive Moderation returns a probabilistic deepfake confidence score rather than a direct binary label. For experimental consistency, predictions above 50% deepfake confidence were classified as FAKE, while predictions below 50% were classified as REAL.

This threshold-based classification enabled standardized accuracy calculation across all evaluated samples.

The detector was evaluated separately on:

- FaceForensics++ real images
- FaceForensics++ deepfake images
- Celeb-DF real images
- Celeb-DF deepfake images
- Real-world online deepfake images

Variables and Measurements

The study examined the performance of a publicly accessible deepfake detection system and human participants across different image datasets. The independent variables were dataset type (FaceForensics++, Celeb-DF, and real-world deepfake images) and image authenticity (real or deepfake).

The dependent variables included detector accuracy, human participant accuracy, real image detection rate, and deepfake detection rate. Detector accuracy was defined as the proportion of correctly classified images based on Hive Moderation's predictions. Human participant accuracy was defined as the proportion of correctly classified images based on participant responses. Real-image detection rate and deepfake detection rate were calculated separately as the

percentage of authentic and manipulated images correctly identified within each dataset.

Data Analysis

Accuracy was calculated as the proportion of correctly classified images divided by the total number of evaluated images. Detector and participant performance were analyzed separately for each dataset. Comparative analysis was performed across datasets to examine changes in performance under increasingly realistic conditions.

The primary evaluation metric used in this study was classification accuracy, calculated as the percentage of correctly classified images.

Additional analysis included:

- Real image detection rate
- Deepfake image detection rate
- Comparison between benchmark and real-world performance
- Comparison between human participants and automated detector outputs

These metrics were used to evaluate the practical reliability of publicly accessible deepfake detection systems under realistic usage conditions.

Ethical Considerations

The human evaluation involved 25 adult participants who voluntarily completed an anonymous image-classification survey. Participants were asked to classify images as either real or deepfake based on visual inspection. No personally identifying information was collected.

RESULTS

Hive Results on FaceForensics++

Hive Moderation was first evaluated using 50 images from the FaceForensics++ dataset, consisting of 25 real images and 25 deepfake images.

Table 2. Hive Performance on FaceForensics++

Category	Correct	Total	Accuracy
Real Images	21	25	84%
Fake Images	18	25	72%
Overall	39	50	78%

Hive correctly classified 21 of 25 real images (84%) and 18 of 25 deepfake images (72%), resulting in an overall accuracy of 78% (39/50 images). Performance was stronger on authentic images than on deepfake images. Several incorrect classifications were associated with high confidence scores despite incorrect predictions.

Table 3. Example Hive Prediction Outputs (FaceForensics++)

Image ID	Ground Truth	Hive Output	Confidence	Correct
FFR01	Real	Deepfake	99.7% Real	No
FFR02	Real	Real	9.7% DF	Yes
FFR03	Real	Deepfake	99.9% DF	No
FFF01	Fake	Real	37% DF	No
FFF02	Fake	Real	63% Real	No

These results indicate that incorrect predictions sometimes occurred with extremely high confidence values.

Hive Results on Celeb-DF

Hive Moderation was further evaluated using 50 images from the Celeb-DF dataset, consisting of 25 real images and 25 deepfake images.

Table 4. Hive Performance on Celeb-DF

Category	Correct	Total	Accuracy
Real Images	21	25	84%
Fake Images	14	25	56%
Overall	35	50	70%

Hive correctly classified 21 of 25 real images (84%) and 14 of 25 deepfake images (56%), resulting in an overall accuracy of 70% (35/50 images). Compared with Face Forensics++, overall accuracy decreased by 8 percentage points (78% to 70%), while deepfake detection accuracy declined from 72% to 56%. Real-image detection accuracy remained unchanged at 84%.

Several Celeb-DF deepfakes were classified as real images or assigned confidence scores below the selected 50% threshold. This pattern is consistent with prior studies suggesting that Celeb-DF contains more visually realistic manipulations and fewer detectable artifacts than earlier benchmark datasets [1, 6].

Human Participant Results

Human participants were evaluated using the same benchmark image sets used during automated detector testing. Participants classified each image as either REAL or FAKE through visual inspection alone.

Table 5. Human Participant Performance

Dataset	Image Type	Correct	Accuracy
FaceForensics++	Real	24/25	96%
FaceForensics++	Fake	24/25	96%
Celeb-DF	Real	15/25	60%
Celeb-DF	Fake	16/25	64%
Real-World Deepfakes	Deepfake Images	25/50	50%

Participants correctly classified 96% of both, real and deepfake images in the FaceForensics++ dataset. Performance declined on Celeb-DF, where participants achieved 60% accuracy on real images and 64% accuracy on deepfake images.

Compared with FaceForensics++, participant accuracy decreased by 36 percentage points for real images (96% to 60%) and 32 percentage points for deepfake images (96% to 64%). These findings suggest that Celeb-DF images were substantially more difficult for human observers to identify through visual inspection alone.

An additional human evaluation was conducted using real-world deepfake images to compare human performance with automated detection under practical conditions.

Hive Results on Real-World Deepfakes

Real-world deepfake images collected from publicly circulated online sources were also evaluated using Hive Moderation.

Table 6. Real-World Deepfake Detection Performance

Dataset Type	Correct	Total	Accuracy
Real-World Deepfakes	25	50	50%

Hive Moderation correctly classified 50% of the evaluated real-world deepfake images. Human participants also achieved 50% accuracy on the same image set.

Compared with benchmark performance, both Hive Moderation and human participants demonstrated substantially lower accuracy on real-world deepfakes. Several real-world samples that appeared convincing to human observers were also incorrectly classified by the automated detector.

Summary of Experimental Results

Table 7. Overall Experimental Summary

Evaluation Category	Accuracy
Hive on FaceForensics++	78%
Hive on Celeb-DF	70%
Human Participants on FaceForensics++	96%
Human Participants on Celeb-DF	62%
Hive on Real-World Deepfakes	50%
Human Participants on Real-World Deepfakes	50%

Across all evaluations, Hive Moderation achieved 78% overall accuracy on FaceForensics++ and 70% overall accuracy on Celeb-DF, with deepfake detection accuracy decreasing from 72% to 56%. Human participants achieved 96% accuracy on both real and deepfake FaceForensics++ images, but performance declined to 60% and 64%, respectively, on Celeb-DF. Both Hive Moderation and human participants achieved 50% accuracy on the real-world deepfake sample set.

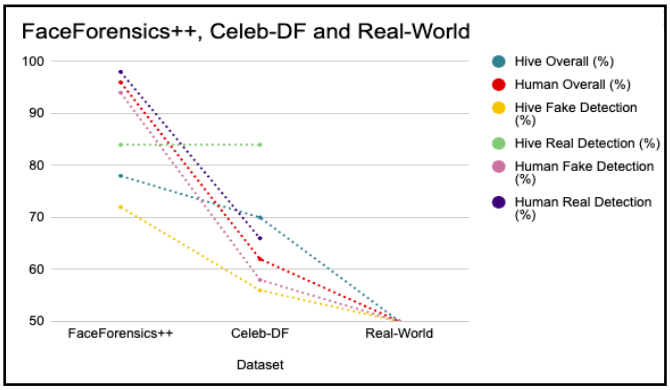

Figure 2. Comparison of Hive Moderation and human detection performance across FaceForensics++, Celeb-DF, and real-world deepfake datasets.

Overall, detector and human performance declined as the evaluated images became more realistic. The highest accuracy was observed on FaceForensics++, followed by Celeb-DF, while the lowest performance was observed on real-world deepfake images.

DISCUSSION

Restatement of Key Findings

The results demonstrated that the performance of both Hive Moderation and human participants varied across datasets. Hive Moderation achieved an overall accuracy of 78% on FaceForensics++, 70% on Celeb-DF, and 50% on the real-world deepfake samples. Human participants achieved 96% accuracy on FaceForensics++, 62% on Celeb-DF, and 50% on the real-world samples. These findings indicate that detection performance declined as the evaluated content became more realistic and representative of real-world online media [3, 4]. The accessibility survey also revealed that relatively few publicly referenced deepfake detection systems were consistently available and usable for public experimentation [2].

Implications and Significance

The findings of this study have implications beyond detector accuracy alone. During the accessibility survey, many publicly referenced deepfake detection systems were found to

be restricted, paid, unavailable, non-functional, or primarily intended for enterprise and investigative use [2]. Rather than weakening the study, this limitation represents an important practical finding. Despite growing public concern about deepfakes and AI-generated misinformation, there remains a significant lack of reliable, free, and easily accessible detection systems for ordinary users. This accessibility gap itself forms part of the study's contribution, as prior research has often focused on detector performance without examining whether such systems are realistically available for public use [2].

These findings suggest that benchmark performance may not fully reflect practical reliability when detectors are applied to more realistic online content [3, 4]. Consequently, the study contributes to the growing body of literature indicating that benchmark accuracy alone should not be considered sufficient evidence of real-world effectiveness.

The accessibility survey also revealed that many publicly available detection tools focused primarily on video-based analysis, while comparatively few reliable image-based detection systems were consistently accessible for public experimentation. This imbalance may reflect fundamental differences between image-based and video-based deepfake detection. Video deepfakes often contain temporal artifacts such as abnormal facial movements, blinking irregularities, lip-synchronization inconsistencies, and frame-to-frame distortions that can provide useful detection signals [1, 6]. In contrast, image-based deepfakes contain only a single frame and therefore lack many of the temporal cues available during video analysis.

This observation may partially explain both the limited availability of image-focused detectors and the reduced performance observed on realistic deepfake images. Given the widespread sharing of manipulated images through social media posts, screenshots, messaging platforms, and reposted content, reliable image-based detection remains an important and comparatively underexplored challenge. By examining both detector accessibility and performance under realistic conditions, this study helps address a gap in the literature between laboratory-based evaluations and the practical experience of ordinary users attempting to verify potentially manipulated media [2].

Recommendations

Future research should evaluate a larger number of publicly accessible deepfake detection systems and include more diverse real-world datasets. Larger participant samples would improve the generalizability of human-comparison results. Additional studies should also investigate video-based deepfake detection under realistic deployment conditions and examine the reliability of detector confidence scores when making classification decisions.

Limitations

Single-detector evaluation: This study primarily evaluated one publicly accessible detector, Hive Moderation, due to the limited availability, accessibility, and consistent functionality of other publicly referenced deepfake detection systems during the evaluation period.

Limited participant sample: The human evaluation involved a relatively small sample of participants. Larger participant studies would improve the statistical strength and generalizability of the findings.

Image-based focus: This study examined only image-based deepfake detection. Full video analysis was not included, meaning the results may not reflect performance when temporal cues such as facial motion, blinking patterns, and frame-to-frame inconsistencies are available.

Limited real-world dataset: The real-world evaluation used a relatively small sample of publicly collected online deepfakes. A larger and more diverse real-world dataset would provide a more comprehensive assessment of detector reliability.

Potential selection bias: The real-world deepfakes included in the study may not fully represent the wide variety of manipulated content currently circulating online.

CONCLUSION

This study evaluated the reliability and accessibility of a publicly available deepfake detection system under both benchmark and real-world conditions. While Hive Moderation demonstrated moderate performance on established benchmark datasets, detection accuracy declined substantially on more realistic deepfake content. Human participants showed a similar pattern, suggesting that increasing realism continues to present challenges for both automated systems and human judgment. In addition, the accessibility survey revealed that many publicly referenced deepfake detection tools were unavailable, restricted, paid, or difficult for ordinary users to access. These findings suggest that benchmark accuracy alone may not provide a complete indication of real-world effectiveness and highlight the need for more reliable, transparent, and publicly accessible detection technologies. As synthetic media becomes increasingly widespread, improving both detector performance and public accessibility will remain important priorities for future research.

ACKNOWLEDGMENTS

I would like to thank all participants who volunteered for the human evaluation component of this study. Their time and contributions were essential to the completion of this research.

I am also grateful to the individuals who provided feedback and encouragement during the development of this project.

REFERENCES

[1] T. Wang, X. Liao, K. P. Chow, X. Lin, Y. Wang. Deepfake detection: A comprehensive survey from the reliability perspective. ACM Computing Surveys. Vol. 57, no. 3, pg. 1–35, 2024. https://doi.org/10.1145/3699932.

[2] M. Rettinger, B. Beaumont, N.-A. Le-Khac, H.-H. Nguyen-Le. How effective are publicly accessible deepfake detection tools? a comparative evaluation of open-source and free-to-use platforms. arXiv. arXiv:2603.04456, 2026.

[3]Y. Lu, T. Ebrahimi. Assessment framework for deepfake detection in real-world situations. EURASIP Journal on Image and Video Processing. Vol. 2024, no. 1, pg. 6, 2024. https://doi.org/10.1186/s13640-024-00691-0.

[4] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, K. Böttinger. Does audio deepfake detection generalize? arXiv preprint arXiv:2203.16263, 2022.

[5] M. Groh, Z. Epstein, C. Firestone, R. Picard. Deepfake detection by human crowds, machines, and machine-informed crowds. Proceedings of the National Academy of Sciences. Vol. 119, no. 1, pg. e2110013119, 2022. https://doi.org/10.1073/pnas.2110013119.

[6] B. C. Soundarya, H. L. Gururaj. Deepfake detection: critical review of state-of-the-art approaches and future perspectives. Discover Applied Sciences. Vol. 8, 2026, https://doi.org/10.1007/s42452-025-08174-9.

[7] V. N. Tran, S. G. Kwon, S. H. Lee, H. S. Le, K. R. Kwon. Generalization of forgery detection with meta deepfake detection model. IEEE Access. Vol. 11, pg. 535–546, 2022. https://doi.org/10.1109/ACCESS.2022.3233969 .

[8] Y. Lai, H. Wang, J. Yang, X. Kang, B. Li, L. Shen, Z. Yu. GM-DF: Generalized multi-scenario deepfake detection. Proceedings of the 33rd ACM International Conference on Multimedia. pg. 4300–4309, 2025, DOI: 10.1145/3746027.3755386.

