American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

Deepfake-Resistant Verification for Public Information Channels

Dr. Neeraj Kulkarni

Assistant Professor, Delhi Technological University (DTU), India

Abstract

Public information channels increasingly distribute emergency announcements, administrative instructions, public-health communication, institutional statements, audiovisual briefings, and other high-consequence information through websites, social platforms, messaging services, livestreams, and mobile communication systems. Generative artificial intelligence has complicated this environment by making realistic synthetic audio, video, images, and impersonation content easier to produce. A technically convincing deepfake can exploit the public's expectation that familiar faces, voices, logos, interfaces, or communication channels indicate authenticity. Consequently, the security objective can no longer be limited to identifying whether media "looks fake." This study proposes a Deepfake-Resistant Public Information Verification Framework (DR-PIVF) that combines source authentication, digital signatures, provenance records, trusted timestamps, cross-channel corroboration, automated forensic detection, and human escalation. NIST's synthetic-content guidance treats provenance, watermarking, labeling, and detection as complementary technical approaches rather than interchangeable solutions, while C2PA provides an open technical standard for recording and verifying the origin and history of digital content.

Because no authentic public-agency verification dataset was supplied, a transparent simulation of 300 hypothetical public-information items was used to compare five verification configurations. The simulated false-acceptance rate declined from 18.7% under visual review alone to 12.9% using deepfake detection alone, 8.4% with source signing and trusted timestamping, 5.1% with provenance plus signature validation, and 2.6% under layered verification. These values are methodological illustrations rather than observed institutional outcomes. The study argues that high-trust public communication should prioritize proof of authorized origin and chain of custody while using deepfake detectors as supplementary forensic evidence. A resilient public-information architecture should also preserve alternative authenticated channels, clear verification indicators, key-management procedures, rapid correction protocols, and human review for conflicting evidence.

Keywords: deepfake verification; public information channels; content provenance; C2PA; synthetic media; digital signatures; information integrity; media authentication; public communication; generative artificial intelligence

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

1. Introduction

Digital public communication increasingly depends on audiovisual media and rapidly distributed electronic messages. Government agencies, universities, hospitals, municipalities, emergency authorities, public-service organizations, and other institutions may communicate through official websites, social-media accounts, livestream platforms, mobile applications, email, SMS, and messaging channels. These systems enable fast dissemination, but their openness also creates opportunities for impersonation and manipulated media. The NSA, FBI, and CISA have warned that synthetic media can be used to impersonate leaders, damage institutional brands, facilitate fraudulent communication, and disseminate false information concerning political, social, military, or economic issues.

Deepfake detection has consequently become an important research area, but detection should not be confused with complete verification. NIST's Open Media Forensics Challenge evaluates manipulation and deepfake-detection systems through performance measures such as area under the receiver-operating-characteristic curve and correct detection at specified false-alarm rates, illustrating that detection is inherently probabilistic rather than an absolute authenticity judgment. Contemporary research also continues to identify generalization to previously unseen manipulation methods and domains as a practical challenge for deepfake detectors. A public institution therefore requires a verification architecture capable of functioning even when a new manipulation technique is not well represented in a detector's training data.

Content provenance offers a different form of evidence. Instead of asking only whether pixels, audio, or video contain forensic artifacts, provenance systems attempt to record where content came from and how it changed. The C2PA technical standard is designed to establish the origin and editing history of digital content through cryptographically verifiable provenance information. Content Credentials based on this standard can expose information concerning creation and modification history when that information is available and verifiable. This approach is particularly relevant to official communication because citizens often need to determine not merely whether a video appears manipulated but whether an authorized public institution actually issued it.

Provenance is nevertheless not a complete solution in isolation. A valid signature proves a relationship between content and a signing credential, but organizational trust still depends on secure key management, legitimate credential issuance, reliable identity binding, and resistance to compromised authorized accounts. Independent 2026 research has also challenged assumptions that current provenance mechanisms should be relied upon as the sole control in high-stakes environments and has demonstrated potential conflicts between provenance and watermarking layers. NIST similarly frames synthetic-content transparency as a combination of authentication, provenance, labeling, watermarking, detection, testing, and auditing rather than a single definitive mechanism.

The present study therefore develops a layered model for deepfake-resistant verification of public information channels. Verification is defined as the process of establishing sufficient evidence that a public-information item originated from an authorized source, has not undergone unauthorized modification, corresponds with the claimed communication context, and can be corroborated through independent institutional channels when necessary. Because no genuine government or public-service dataset was supplied, the analysis is explicitly simulation-based. The objective is not to claim that a particular technology eliminates deepfakes, but to demonstrate how public institutions can combine

cryptographic provenance, source authentication, trusted publication infrastructure, forensic detection, cross-channel confirmation, and human review into a more defensible verification system.

2. Objectives of the Study

2.1 To Develop a Layered Verification Architecture

The first objective is to develop a security architecture that moves beyond detector-only deepfake screening. The proposed model integrates authorized-source identity, digital signing, timestamp validation, content provenance, cross-channel corroboration, forensic analysis, and structured human escalation. This follows NIST's position that provenance and detection are complementary components of synthetic-content transparency.

2.2 To Compare Alternative Verification Approaches

The second objective is to compare, through simulation, the false-acceptance risk associated with visual review, automated deepfake detection, signed-source verification, provenance-enabled verification, and a fully layered verification workflow. False acceptance refers to a manipulated or unauthorized item being incorrectly treated as authentic.

2.3 To Identify Governance Requirements for Public Information Integrity

The third objective is to identify institutional requirements necessary for resilient public communication, including signing-key governance, authorized publisher registries, provenance preservation, trusted timestamps, redundant authenticated channels, incident escalation, public correction mechanisms, and verification interfaces understandable to nontechnical users.

3. Review of Literature

3.1 Deepfake Detection and the Limits of Forensic Classification

Deepfake detection research has progressed significantly since early large-scale facial-forgery datasets enabled systematic evaluation. Rössler and colleagues' FaceForensics work created a large dataset for studying automated facial manipulation and provided an important foundation for subsequent detection research. Since then, research has expanded from specific manipulation types toward models intended to generalize across previously unseen generators, compression conditions, and operational environments.

The continuing research emphasis on generalization demonstrates why detector outputs should be interpreted probabilistically. Recent work explicitly describes unseen manipulation techniques as a continuing practical deployment challenge, and multi-scenario research has found that combining heterogeneous forgery datasets can itself create domain discrepancies that reduce straightforward model performance. Research published in 2026 continues to investigate suppression of manipulation-specific shortcuts because detectors may learn artifacts associated with known generation methods rather than stable properties that transfer reliably to unknown attacks.

NIST's OpenMFC program reflects this evaluation-oriented perspective by testing manipulation detection, image deepfake detection, and video deepfake detection under controlled conditions and by requiring confidence scores rather than binary claims of certainty. For public-information verification, this distinction is critical. A detector can provide evidence that media resembles known synthetic

content, but a low manipulation score does not independently prove that the source is an authorized public institution.

3.2 Provenance, Content Credentials, and Source Authentication

NIST AI 100-4 provides a broader framework for synthetic-content transparency by examining authentication and provenance, watermarking, labeling, detection, testing, auditing, and maintenance. Authentication approaches can provide affirmative evidence concerning content origin rather than relying entirely on forensic indications of manipulation.

C2PA is a prominent implementation of this provenance-oriented approach. The coalition develops an open technical standard intended to establish the source and history of digital content. Its Content Credentials model uses cryptographically bound provenance structures so that supported tools can inspect information concerning the content's history and edits. The Content Authenticity Initiative provides verification tooling that can inspect such credentials when they are present.

For public information, provenance can support a transition from “detecting fake media” toward “verifying official media.” An authorized agency could sign content at publication, bind that content to an institutional credential, include a trusted timestamp, and make verification available through an official endpoint. A copied or manipulated item could then be compared with the authenticated original.
However, provenance systems remain dependent on trustworthy implementation. Independent 2020 research has argued that C2PA should not be treated as sufficient by itself for high-stakes journalism, legal evidence, or similar contexts because system-level trust assumptions and protocol weaknesses may remain. Separate research has demonstrated that cryptographic provenance and watermarking systems can produce conflicting authenticity signals when evaluated independently, motivating cross-layer verification rather than isolated checks. These findings reinforce the need for layered evidence.

3.3 Verification Governance for High-Trust Public Communication

The organizational response to deepfakes must extend beyond technical media analysis. NSA, FBI, and CISA guidance recommends that organizations prepare for synthetic-media threats through verification capabilities, provenance-related technologies, information sharing, rehearsed response procedures, and protection of high-priority communications. For public agencies, these recommendations imply that authentication must be designed into the communication process before a deepfake incident occurs.

Public channels also require redundancy. If a suspicious video claims to contain an emergency statement, users should be able to confirm the same message through an independently authenticated website, alert system, or institutional registry. This principle is particularly important when an adversary compromises or convincingly imitates one distribution channel. The present study calls this cross-channel corroboration.

The literature therefore reveals a shift from content-only detection toward broader media-integrity systems. NIST examines authentication, provenance, watermarking, and detection as distinct but complementary approaches; C2PA provides a provenance standard; and federal cybersecurity guidance recommends organizational preparation and real-time verification capabilities. The remaining governance challenge is to combine these capabilities into a public-facing verification model that is technically robust, operationally realistic, and understandable to citizens.

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

4. Research Methodology

4.1 Research Design

The study adopts a simulation-based comparative research design. No government agency, emergency authority, public broadcaster, or public institution provided real verification records. No authentic deepfake incident dataset was used to generate the reported numerical results.

A synthetic dataset representing 300 hypothetical public-information items was constructed. These items conceptually include video briefings, audio statements, emergency notices, public-service announcements, and institutional social-media communications.

Five verification conditions were modeled: visual review only; automated deepfake-detector screening; digital signature with trusted source identity and timestamp; provenance plus cryptographic signature; and layered verification combining provenance, signature validation, source identity, timestamp verification, cross-channel corroboration, forensic detection, and human escalation.

The principal outcome was the false-acceptance rate, representing the percentage of manipulated or unauthorized items hypothetically accepted as authentic. Lower values indicate stronger verification performance.

4.2 Proposed Verification Measurement Instrument

A future empirical investigation should evaluate both technical performance and user confidence. The following five questionnaire items are proposed for administrators, public-information officers, journalists, emergency personnel, or other users responsible for validating institutional communications.

Table 1: Proposed Five-Item Questionnaire for Public Information Verification

Item	Proposed Questionnaire Statement	Construct
Q1	I can determine whether a public-information item originated from an officially authorized publishing source.	Source authenticity
Q2	The verification interface clearly indicates whether the content's digital signature and provenance record are valid.	Cryptographic verification clarity
Q3	I can distinguish confirmed information from content that has missing, conflicting, or inconclusive authenticity evidence.	Uncertainty recognition
Q4	Important public announcements can be independently confirmed through another trusted institutional channel.	Cross-channel corroboration
Q5	I know how and where to escalate suspicious public media when automated verification produces uncertain or conflicting results.	Human escalation readiness

Source: Proposed by the author for future empirical validation. These questions were not administered in the present simulation.

The instrument should be validated before empirical use through reliability analysis, content validation, construct validation, and comparison with observed verification behavior. A future field study should avoid interpreting user confidence as equivalent to technical authenticity because highly confident users can still accept manipulated content.

4.3 Analytical Procedure and Ethical Safeguards

The synthetic analysis compares false-acceptance rates across the five verification approaches. The modeled values were chosen to illustrate a progressively layered security architecture rather than to estimate real institutional performance.

5. Across the five simulated conditions, verification maturity and false acceptance produced a descriptive correlation of approximately r = −0.

987. This value reflects the deliberately structured simulation and must not be generalized as a real-world effect size.

The research design also distinguishes absence of provenance from evidence of falsity. Content may lack credentials because the producing tool does not support a provenance standard, metadata was removed, or the content originated through an older workflow. The absence of a provenance record should therefore trigger additional verification rather than automatic classification as fake. Content Credentials themselves are still not universally present across digital media.

No human-participant ethical approval was required for the synthetic analysis. A future empirical study should nevertheless protect confidential signing infrastructure, avoid publishing exploitable key-management weaknesses, prevent exposure of sensitive emergency-communication processes, and distinguish legitimate security testing from deceptive public dissemination.

5. Analysis

5.1 Comparative Performance of Verification Approaches

The simulation demonstrates progressive risk reduction as the verification architecture becomes more layered. Visual review alone produced a simulated false-acceptance rate of 18.7%. This condition represents dependence primarily on human perception, familiar faces, visual appearance, logos, and communication style.

Automated deepfake detection reduced the simulated rate to 12.9%. This improvement reflects the value of forensic analysis but also preserves substantial residual risk because detectors do not constitute affirmative source authentication and may encounter previously unseen manipulation methods. Current research continues to identify cross-domain and unseen-manipulation generalization as an active challenge.

Adding authenticated source identity and trusted timestamping reduced the simulated false-acceptance rate to 8.4%, while provenance combined with cryptographic signing reduced it to 5.1%. The strongest condition—layered verification—produced a simulated false-acceptance rate of 2.6%.

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

Table 2: Simulated Verification Outcomes

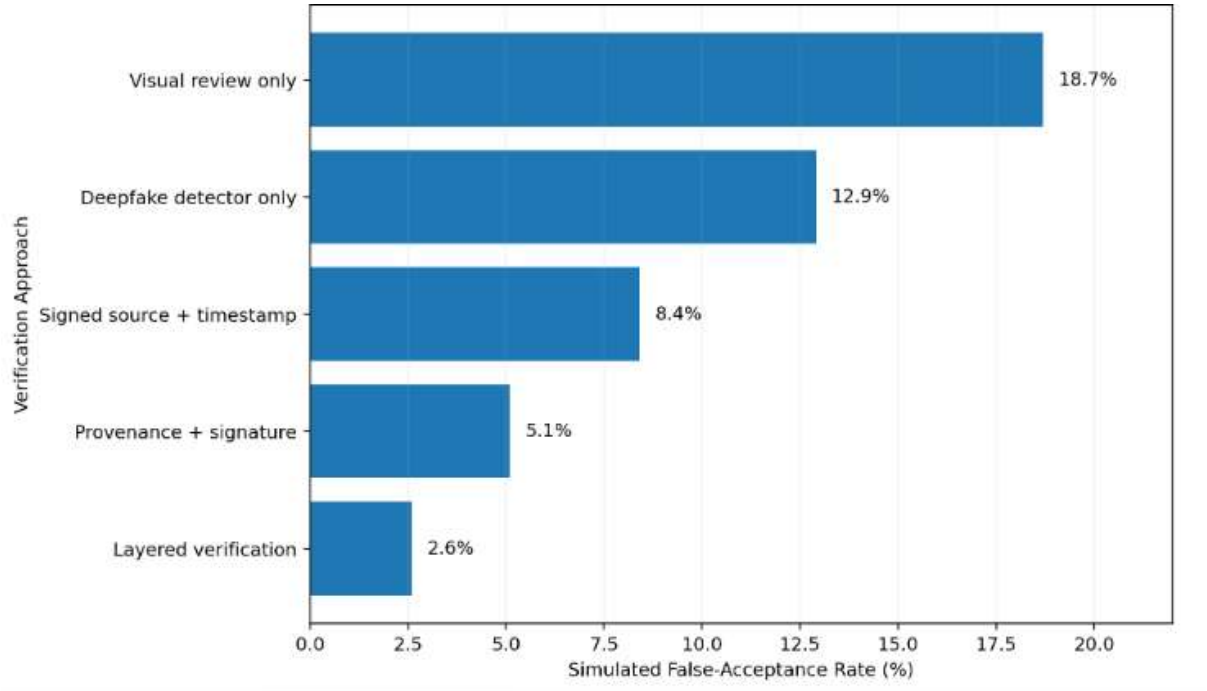
Verification Approach	Simulated Items	False-Acceptance Rate	Correct Authentication Rate	Mean Verification Time
Visual Review Only	60	18.7%	76.8%	92 sec
Deepfake Detector Only	60	12.9%	82.6%	54 sec
Signed Source + Trusted Timestamp	60	8.4%	88.5%	36 sec
Provenance + Cryptographic Signature	60	5.1%	92.1%	28 sec
Layered Verification	60	2.6%	95.3%	34 sec

Source: Author-generated synthetic dataset; total simulated n = 300. None of these values represent measured government or platform performance.

The slight increase in verification time from the provenance-only condition to layered verification represents additional cross-channel and escalation checks. Within the simulation, this tradeoff produces substantially stronger authenticity assurance.

5.2 False-Acceptance Risk Across Verification Levels

Graph 1: Simulated False-Acceptance Risk Across Public-Information Verification Approaches



Source: Author-generated synthetic analysis.

Interpretation: The graph demonstrates that simulated false-acceptance risk decreases as verification moves from appearance-based judgment and detector-only screening toward authenticated provenance and layered corroboration.

Key Finding: The strongest simulated performance does not arise from a more accurate deepfake detector alone. It arises from combining multiple evidence types that answer different questions: whether the media appears manipulated, whether it came from an authorized issuer, whether its history is intact, and whether the message can be independently corroborated.

5.3 Verification Logic for Public Information

The analytical pattern supports a distinction between detection confidence and authenticity assurance. Detection confidence represents an estimate that a piece of content contains manipulation characteristics. Authenticity assurance represents the broader evidentiary basis for believing that the communication was issued by the claimed organization and has retained an acceptable chain of integrity.

NIST's synthetic-content framework provides support for this distinction by treating authentication and provenance separately from synthetic-content detection. A detector can flag an apparently manipulated item even when provenance evidence exists, while a valid provenance record can coexist with complex trust problems if credentials, toolchains, or assertions are compromised or interpreted incorrectly. Independent 2026 analyses further show why contradictory verification signals require joint evaluation.

A high-trust public-information system should therefore classify results into states such as verified, unverified, conflicting evidence, or verification unavailable rather than presenting every item as simply “real” or “fake.” This model preserves uncertainty and makes escalation possible.

6. Discussion

6.1 Verification Should Prioritize Authorized Origin Rather Than Visual Realism

The principal finding of the study is that deepfake-resistant communication requires a change in the underlying security question. Traditional media-forensics systems ask whether an item has characteristics associated with manipulation. Public communication additionally requires an affirmative answer to a different question: Did the authorized institution issue this information?

This distinction becomes increasingly important as synthetic-media quality improves. NIST's deepfake evaluation activities demonstrate that detection systems operate through confidence scores and measurable error rates rather than absolute certainty. Recent detector research likewise continues to improve cross-manipulation and cross-domain generalization, confirming that unseen attacks remain a relevant research problem. An authentication architecture should therefore remain functional even if a detector is temporarily unable to recognize a novel synthesis technique.

Cryptographic source authentication provides a stronger foundation because it changes the verification task from visual inference to evidence of authorized publication. C2PA similarly focuses on source and content history rather than merely estimating whether media contains synthetic artifacts. For public institutions, this suggests that authenticated publishing pipelines should become part of communication infrastructure rather than a post-publication forensic service.

The public-facing interface must nevertheless remain understandable. Citizens should not need to interpret certificate chains, cryptographic hashes, or forensic-model probabilities. A verification interface can instead communicate whether the institutional identity was validated, whether the content has an intact provenance record, whether important edits were recorded, and whether an independent official channel contains the same message.

6.2 Provenance and Detection Should Be Treated as Complementary Controls

The second major implication concerns defense in depth. NIST AI 100-4 does not position provenance, watermarking, labeling, and detection as mutually exclusive choices; instead, it analyzes them as components of a broader technical transparency environment. The simulated findings adopt the same logic.

Provenance is strongest when trusted information is attached at creation or publication and remains verifiable throughout distribution. Detection is useful when provenance is absent, stripped, damaged, unsupported, or suspicious. Cross-channel verification can provide additional assurance when provenance and detector outputs disagree. Human experts remain necessary when automated evidence is ambiguous or when the consequences of a false decision are unusually high.

Current research also demonstrates why this layered approach is necessary. A 2020 independent security analysis questions whether present C2PA mechanisms alone are sufficient for high-stakes trust decisions, while other researchers have demonstrated cases in which valid provenance and watermark signals can contradict one another. These studies do not make provenance useless; rather, they show that provenance should be integrated into a broader verification architecture with explicit conflict handling.

A robust system should therefore avoid a single green check mark derived from one control. Verification should aggregate multiple independent signals and expose uncertainty when the evidence does not converge.

6.3 Institutional Governance Is as Important as Verification Technology

Deepfake-resistant verification ultimately depends on organizational governance. A perfect signing algorithm cannot protect a public-information system if an authorized private key is stolen, if departed staff retain publishing access, if malicious insiders can sign unauthorized material, or if citizens cannot identify the legitimate verification endpoint.

Institutions therefore require formal ownership of signing credentials, key rotation, revocation procedures, access logging, multifactor administrative controls, separation of publishing authority, emergency key replacement, and independent audit. They also require operational procedures for situations in which false audiovisual content is already circulating. NSA, FBI, and CISA guidance emphasizes planning, real-time verification capability, personnel preparation, and coordinated response to synthetic-media threats.

Redundancy is equally important. An emergency announcement should not depend on a single social-media account whose compromise could temporarily make authentic and fraudulent communications indistinguishable to the public. Critical messages should be reproducible across independently controlled authenticated channels. A citizen who sees a controversial video should be able

to verify the underlying message through an official website, authenticated notification service, or public verification portal.

The broader governance objective is therefore verifiable institutional communication, not merely deepfake detection. A resilient public-information ecosystem should make authentic information easy to confirm, suspicious information easy to escalate, and uncertainty visible rather than concealed.

7. Conclusion

The rapid development of synthetic-media technologies has altered the trust assumptions underlying public digital communication. Familiar appearance, recognizable voice, official logos, and professional video quality can no longer provide sufficient evidence that an announcement is authentic. Deepfake detectors remain valuable, but their outputs are probabilistic and current research continues to investigate robustness and generalization to new manipulation techniques. Public institutions consequently require mechanisms that establish authorized origin in addition to mechanisms that search for signs of manipulation.

The present study proposes a layered Deepfake-Resistant Public Information Verification Framework integrating authenticated source identity, digital signatures, trusted timestamps, provenance, cross-channel corroboration, forensic detection, and human escalation. This architecture is consistent with NIST's broader treatment of authentication, provenance, watermarking, and detection as complementary synthetic-content safeguards and with C2PA's effort to establish verifiable origin and history for digital content. At the same time, recent independent research demonstrates that provenance mechanisms should not be treated as infallible or used without governance and cross-layer validation.

The simulation of 300 hypothetical public-information items illustrates the proposed analytical relationship. The false-acceptance rate declines from 18.7% under visual review alone to 12.9% under detector-only screening and ultimately to 2.6% within the fully layered verification condition. These results are deliberately synthetic and should not be interpreted as observed performance of C2PA, government communication systems, deepfake detectors, or any real institution. Their purpose is to demonstrate a testable evaluation framework and to show how progressively independent evidence sources can be compared.

Future research should validate the framework through controlled experiments involving authenticated institutional communications, multiple media types, real signing infrastructure, diverse deepfake generators, provenance-preserving and provenance-stripping transformations, adversarial channel compromise, and realistic public users. Particular attention should be given to false reassurance, missing credentials, compromised signing keys, accessibility of verification interfaces, key revocation, cross-platform provenance preservation, and contradictory detector/provenance results. Deepfake resistance should ultimately be evaluated not by whether a single tool can label content as fake, but by whether citizens and institutions can reliably establish who issued a message, what happened to it after publication, whether independent evidence agrees, and what action should be taken when authenticity remains uncertain.

References

1. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). DeepFakes and beyond: A survey of face manipulation and fake detection. *Information Fusion, 64*, 131–148. https://doi.org/10.1016/j.inffus.2020.06.014
2. Yu, P., Xia, Z., Fei, J., & Lu, Y. (2021). A survey on deepfake video detection. *IET Biometrics, 10*(6), 607–624. https://doi.org/10.1049/bme2.12031
3. Pasquini, C., Amerini, I., & Boato, G. (2021). Media forensics on social media platforms: A survey. *EURASIP Journal on Information Security, 2021*, 4. https://doi.org/10.1186/s13635-021-00117-2
4. Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. (2020). On the detection of digital face manipulation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5781–5790.
5. Jiang, L., Li, R., Wu, W., Qian, C., & Loy, C. C. (2020). DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2889–2898.
6. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1–11.
7. Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2020). The DeepFake Detection Challenge dataset. *arXiv preprint arXiv:2006.07397*.
8. Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A new dataset for deepfake forensics. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 390–399.
9. Li, Y., Chang, M.-C., & Lyu, S. (2018). In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking. *IEEE International Workshop on Information Forensics and Security*, 1–7.
10. Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A compact facial video forgery detection network. *IEEE International Workshop on Information Forensics and Security*, 1–7.
11. Nguyen, H. H., Fang, F., Yamagishi, J., & Echizen, I. (2019). Multi-task learning for detecting and segmenting manipulated facial images and videos. *IEEE International Conference on Biometrics Theory, Applications and Systems*, 1–8.
12. Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Capsule-forensics: Using capsule networks to detect forged images and videos. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2307–2311.
13. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020). Face X-ray for more general face forgery detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5001–5010.
14. Wang, S.-Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). CNN-generated images are surprisingly easy to spot... for now. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8695–8704.
15. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., & Holz, T. (2020). Leveraging frequency analysis for deep fake image recognition. *International Conference on Machine Learning*, 3247–3258.

16. Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019). Protecting world leaders against deep fakes. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 38–45.
17. Tursman, E., George, M., Kamara, S., & Tompkin, J. (2020). Towards untrusted social video verification to combat deepfakes via face geometry consistency. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 654–655.
18. Verdoliva, L. (2020). Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing, 14*(5), 910–932.
19. Masood, M., Nawaz, M., Malik, K. M., Javed, A., & Irtaza, A. (2021). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *arXiv preprint arXiv:2103.00484*.
20. Cao, X., & Gong, N. Z. (2021). Understanding the security of deepfake detection. *arXiv preprint arXiv:2107.02045*.