West Science Law and Human Rights
Vol. 4, No. 03, July 2026, pp. 299~309

Reconstructing Reliability Standards for Deepfake Detectors as Electronic Evidence in Indonesia's Criminal Justice System

Muhammad Abdul Azis[1], Pujiyono[2], Riski Ananda Kusuma Putri[3]
[1,2] Universitas Diponegoro
[3] Universitas Muhammadiyah Cirebon

Article Info

Article history:

Received Jul, 2026
Revised Jul, 2026
Accepted Jul, 2026

Keywords:

Artificial Intelligence
Criminal Procedure
Deepfake Detector
Electronic Evidence
Forensic Reliability

ABSTRACT

The development of deepfakes has transformed the evidentiary problem of video and audio from file authentication into an assessment of the computational method used to determine media authenticity. This study has two objectives: to examine the legal position and limits of the probative value of deepfake detector outputs under Indonesian criminal procedure and to reconstruct an AI Forensic Reliability Test that safeguards authenticity, scientific reliability, and procedural fairness. It applies normative legal research through statutory, conceptual, comparative, and interdisciplinary approaches. The first finding indicates that detector outputs may enter proceedings through electronic evidence, examination reports, and expert testimony, but they do not constitute independent evidence or binary statements of truth. Their value depends on object authenticity, lawful acquisition, methodological validity, error rates, and corroboration. The second finding develops an AI Forensic Reliability Test comprising eight dimensions: object authenticity, model validity, data quality, error rates and calibration, robustness and generalization, traceability and reproducibility, expert competence and independent testing, and procedural fairness and accountability. The framework operationalizes the negative statutory theory of proof, epistemic reliability, machine testimony, due process of law, equality of arms, and accountable algorithms.

Corresponding Author:

Name: Muhammad Abdul Azis
Institution Address: Universitas Diponegoro
e-mail: muhammadabdulazis@students.undip.ac.id

1. INTRODUCTION

Generative models can produce synthetic faces, lip movements, and voices that resemble specific individuals through techniques such as face swapping, facial reenactment, lip synchronization, and synthetic speech [1], [2]. These capabilities reduce the intuitive reliability of audiovisual recordings as direct representations of events. The examination of such evidence must therefore distinguish between file integrity, content authenticity, and the reliability of the detection method employed.

Deepfakes may be used for fraud, extortion, image-based sexual abuse, impersonation, and disinformation [3], [4]. This technology also enables perpetrators to deny the authenticity of genuine recordings by claiming that they have been manipulated,

a phenomenon known as the liar's dividend [3], [5]. Its consequences include disruption to investigations, credibility assessments, evidence authentication, and public trust in audiovisual information [4], [6].

Deepfake detectors operate by identifying spatial, temporal, frequency-based, physiological, or audiovisual inconsistencies within synthetic media [7], [8]. Their outputs generally take the form of classification labels or scores that do not directly establish the identity of the creator, the purpose of the content, or the defendant's involvement [7], [9]. Detector accuracy may also decline when the system encounters new types of generators, platform compression, changes in resolution, video cropping, or data that differ from the training dataset [9], [10]. False positives may direct an investigation toward the wrong person, while false negatives may cause manipulated content to be accepted as an authentic recording.

A detector score cannot be equated with a legal fact because it results from statistical relationships between input data and machine-learning parameters [11], [12]. A score of 0.92, for example, does not necessarily indicate a 92 percent probability that a recording was falsified by the defendant. It may merely reflect the system's level of confidence in assigning the recording to a particular category [9], [12]. Its use in evidentiary proceedings requires an examination of the detector's knowledge base, operating conditions, error rate, and the parties' opportunity to challenge the process through which the conclusion was generated [11], [12].

Law Number 20 of 2025 concerning the Criminal Procedure Code, which entered into force on 2 January 2026, recognizes electronic evidence as a legally admissible form of evidence under Article 235 paragraph (1) [13]. Article 235 paragraphs (3) to (5) require authentication and lawful acquisition and provide for the exclusion of evidence that is unauthentic or unlawfully obtained. Article 242 covers electronic information, electronic documents, and electronic systems related to a criminal offence [13]. These provisions correspond with Article 5 of the Electronic Information and Transactions Law, which recognizes electronic information, electronic documents, and their printed outputs as legally valid evidence [14].

This normative framework has not established specific parameters for assessing the reliability of deepfake detectors. Article 238 of the Criminal Procedure Code regulates expert testimony but does not prescribe the minimum technical information that must be disclosed when an expert opinion relies on machine learning [13]. This regulatory gap may result in detection outputs being accepted without adequate methodological examination or rejected without sufficient scientific justification.

Previous studies have examined deepfakes from the perspectives of privacy, evidence, regulation, and technical validity. Kusnadi and Putri found that identity manipulation through deepfakes reduces victims' control over their personal data and requires stronger legal protection [15]. Antoro, Budhijanto, and Somawijaya argued that artificial intelligence outputs may be considered electronic evidence when they satisfy the requirements of authenticity, relevance, integrity, and expert examination [16]. Van der Sloot and Wagensveld demonstrated that synthetic media blur the distinction between authentic and manipulated information, requiring regulation that protects privacy, freedom of expression, and public trust [17]. Hydara, Kikuchi, and Ozono found that detectors may assist in assessing media authenticity, although their results are affected by the type of manipulation, data characteristics, and testing conditions [18].

These studies have not specifically addressed the application of reliability standards to deepfake detectors in Indonesian criminal proceedings, particularly the relationship between technical accuracy, evidence authenticity, expert examination, and judicial assessment. This gap creates uncertainty regarding the permissible scope of detection results and the legal consequences of relying on an unvalidated examination. This study analyses the application of reliability standards to

deepfake detectors and their implications for evidentiary weight, the protection of defendants' rights, and the prevention of erroneous judgments.

The novelty of this study lies in its examination of how deepfake detector reliability standards should be applied when determining the admissibility, evidentiary weight, or exclusion of electronic evidence. Their application is expected to reduce reliance on unverified detection scores and assist judges in evaluating evidence transparently and accountably.

This study is important because the increasing use of digital media as evidence is accompanied by the growing capacity of deepfake technology to produce visual and audio manipulations that are difficult to identify. In the absence of reliability standards, falsified content may be accepted as authentic evidence, while genuine recordings may be rejected because of detection errors. Such conditions may prejudice victims, restrict defendants' right to challenge evidence, and undermine the objectivity of judicial decisions.

This study aims to analyse the legal status and limitations of the probative value of deepfake detector outputs under Indonesian criminal procedural law. It also seeks to formulate a reliability standard for assessing authenticity, technical validity, and procedural fairness in the examination of electronic evidence.

2. METHODS

This study employs a normative legal research method by examining evidentiary rules, the concept of reliability, and standards for assessing the outputs of computational systems. The research materials include legislation, legal principles, doctrines, judicial decisions, and scholarly literature used to identify normative gaps and formulate legal prescriptions [19], [20].

The statutory approach is used to examine Law Number 20 of 2025 concerning the Criminal Procedure Code [13], Law Number 1 of 2024 concerning the Second Amendment to the Electronic Information and Transactions Law [14], and Law Number 27 of 2022 concerning Personal Data Protection [21]. The conceptual approach is applied to analyse authenticity, admissibility, evidentiary weight, epistemic reliability, machine testimony, due process of law, equality of arms, and accountable algorithms. The comparative approach is used to identify gatekeeping functions, methodological transparency, and adversarial testing mechanisms relevant to the Indonesian criminal evidence system.

An interdisciplinary approach is also applied to understand the technical aspects of deepfake detectors, including datasets, data leakage, calibration, robustness, generalisation, and adversarial attacks [9], [22]. All research materials are analysed qualitatively through the inventory of legal norms, the mapping of technical aspects onto evidentiary requirements, gap analysis, and the formulation of a reliability standard. The analytical results are assessed based on their capacity to ensure evidentiary reliability, protect the rights of the parties, and support practical implementation by law enforcement authorities.

3. RESULTS AND DISCUSSION

3.1 The Legal Status and Limitations of the Probative Value of Deepfake Detector Outputs in the Indonesian Criminal Procedure System

A deepfake detector operates by classifying numerical representations of images, sequences of frames, audio recordings, or audiovisual relationships based on learned patterns rather than directly identifying falsity. Detection methods may analyse spatial artefacts, temporal irregularities, frequency characteristics, inconsistencies in lighting, blinking patterns, lip movements, or traces of synthetic generation [1], [7]. The resulting classification is influenced by class definitions, the sources and labelling of data, system architecture, preprocessing procedures, and testing conditions [8], [9]. Detector outputs must therefore be understood as technically limited

inferences rather than determinations of legal fact.

Detector reliability cannot be adequately assessed through aggregate accuracy alone. False positives may cause authentic media to be classified as falsified and affect an alibi, the direction of an investigation, or the credibility of a witness. False negatives may cause manipulated content to be accepted as genuine documentation of an event [8], [9]. Predictive values also vary according to the prevalence of deepfakes among the media being examined. High specificity does not necessarily eliminate the risk of error [9], [12]. An expert must explain the confusion matrix, false-positive rate, false-negative rate, decision threshold, uncertainty interval, and the reasons for selecting a particular threshold so that performance figures are not used without appropriate context [9], [12].

Generalisation constitutes a major concern because a detector tested on particular generators, resolutions, languages, facial characteristics, and compression formats may not remain accurate under the different conditions of an individual case [9], [10], [23]. Distribution shift, data leakage, and improper validation may produce misleading performance when the system merely recognises dataset-specific characteristics rather than stable indicators of manipulation [22]. The history, composition, labelling, and partitioning of datasets should form part of the basis of an expert opinion. Documentation such as datasheets may clarify the intended purpose, limitations of use, data representation, and risks relevant to the examination of evidence [24].

Reliability is also affected by the history of the file. Cropping, blurring, resizing, noise, transcoding, screen recording, and adversarial attacks may alter predictions or remove the artefacts upon which the detector relies [9], [10]. Testing must represent the condition of the file from its creation, transmission, and storage to its seizure and forensic examination. The examined copy must also be compared with the available source file [7], [9]. Testing only a converted file may conflate preprocessing errors with conclusions concerning the authenticity of its content.

Experimental studies demonstrate that humans and detection systems exhibit different patterns of error, while their combination may improve performance under certain conditions [25]. This finding positions the detector as an instrument assisting the expert rather than as a substitute for forensic examination and judicial assessment. Human oversight has little meaning when an expert merely repeats a system-generated label without understanding the method, operating conditions, and potential errors. Trust in audiovisual recordings must not be blindly transferred to algorithmic outputs because both deepfakes and their detection systems require epistemic examination [5].

The legal status of a detector output must be distinguished from that of the media being examined. Video recordings, audio recordings, metadata, system reports, and printed outputs may qualify as electronic evidence when they relate to a criminal offence and satisfy the applicable legal requirements [13], [14]. A detector output constitutes the result of computational analysis and requires an expert explanation of the method, assumptions, conditions of use, and limits of inference [11], [16]. This distinction prevents the conflation of physical or electronic evidence, examination instruments, analytical reports, and expert opinions.

Article 235 paragraphs (3) and (4) of the Criminal Procedure Code require lawful acquisition and authentication, the assessment of which rests with the court [13]. Authentication of digital media includes verification of the file's identity, integrity, origin, and connection to the relevant event. Hash values, metadata, seizure records, and chain-of-custody documentation may establish that the file examined is identical to the file obtained. Such evidence does not necessarily prove that the content had not been manipulated before seizure. A detector may assist in addressing this issue only to the extent that its method and conditions of use are reliable.

Article 235 paragraph (5) of the Criminal Procedure Code provides that evidence that is unauthentic or unlawfully obtained may not be used and has no probative value [13]. This legal consequence must be distinguished from methodological weaknesses in a detector. A breach of the chain of custody concerns the authenticity of the evidentiary object, while limitations relating to generalisation, calibration, or validation primarily affect the weight of the expert opinion. Detection results may be considered unsuitable for judicial consideration when the method cannot be tested, is incompatible with the characteristics of the evidence, or conceals a material error rate.

Article 238 of the Criminal Procedure Code positions an expert as a person who provides sworn testimony based on specialised knowledge [13]. A deepfake expert should explain their qualifications, acquisition procedures, the device and software version used, system parameters, preprocessing procedures, validation data, performance metrics, test results, limitations, and plausible alternative explanations. The court retains the authority to assess the expert opinion and may not delegate its adjudicative function to a technological system. Detection results may not constitute the sole basis for conviction because Article 244 of the Criminal Procedure Code requires the criminal offence to be established lawfully and beyond the judge's reasonable conviction [13]. A classification score may assist in identifying indications of manipulation, but it does not prove the identity of the creator, the purpose of the content, intent, or the defendant's criminal responsibility [26].

The function of detection results differs at each stage of criminal proceedings. During an investigation, a detector may be used for triage and to determine the need for further examination, but it should not serve as the sole basis for coercive measures. At the evidentiary stage, its output may support an expert opinion when accompanied by the source file, process documentation, and comparative testing. During judicial assessment, the result must be evaluated together with other evidence without predetermined probative force. Admissibility concerns relevance, legality, authenticity, and the adequacy of the methodological foundation. Weight of evidence concerns the strength with which a result supports a particular proposition after considering error rates, the quality of the examined object, consistency with other evidence, and the results of adversarial examination [16], [27].

The processing of faces, voices, and biometric characteristics is also subject to personal data protection law. Biometric data constitute specific personal data whose processing requires a lawful basis, purpose limitation, appropriate security measures, and accountability [21]. A risk of infringement may arise when sensitive files involving victims, witnesses, or third parties are uploaded to cloud-based detection services without adequate safeguards [15]. Technical reliability must be accompanied by appropriate data governance to ensure that the examination of evidence does not create additional harm.

Methodological transparency frequently conflicts with claims of trade secrecy asserted by technology providers. Such claims may restrict a defendant's ability to examine technology that affects the evidence used against them [28]. Trade secret protection need not be abolished, but it may be reconciled with procedural rights through restricted access for independent experts, protective orders, validation reports, system logs, audit interfaces, and the disclosure of minimum information necessary for meaningful examination. Without such access, informational inequality may arise alongside the risk of automation bias, under which technological outputs are treated as objective without sufficient critical scrutiny [29].

The negative statutory theory of proof requires the concurrence of legally recognised evidence and the judge's rationally formed conviction [13], [26]. A detector output does not create a new category of evidence. It enters the proceedings through electronic evidence, an examination report, or expert testimony. A detector

threshold may not be transformed into a threshold for conviction because media classification differs from proving every element of a criminal offence. Detection results must consequently be corroborated through metadata, frame analysis, source-device examination, distribution history, communications, witness testimony, and other evidence. The judgment must also explain the proposition supported by the result, the basis for trusting the method, the identified limitations, and the relationship between the detection output and other evidence. Material methodological uncertainty that cannot be resolved must be assessed according to the principle of in dubio pro reo.

Epistemic reliability requires an assessment that is specific to the conditions of use. A system that performs reliably on a benchmark dataset may not be reliable when applied to case-specific video that has undergone repeated compression. The concept of machine testimony describes system-generated information accepted without direct observation of the system's internal processes [11]. A machine cannot take an oath or assume responsibility for normative conclusions. The expert must therefore explain the process, examine the suitability of the operating conditions, and assume professional responsibility for the interpretation. The selection of a threshold must also consider the purpose of the examination and the different consequences of false positives and false negatives [12].

Due process of law and equality of arms require notice of the use of a detector, access to forensic copies, methodological information, complete reports, parameters, system logs, an opportunity to present a rebuttal expert, and sufficient time to conduct replication. When independent examination is impossible, the evidentiary weight of the result should be reduced or the result should be excluded [28]. Adversarial testing may examine whether the system is suitable for the generator allegedly used, its performance after compression, changes resulting from threshold adjustments, the false-positive rate under comparable conditions, and the use of alternative methods. This mechanism limits expert overclaiming without requiring every party to possess advanced expertise in machine-learning engineering.

The judge functions as both gatekeeper and final assessor, determining the adequacy of the methodological basis and the weight of the detection result. Explainable artificial intelligence may assist in explaining a process, but visual explanations such as heatmaps do not automatically prove that a particular area has been manipulated without additional validation [30]. Accountability requires the identification of the parties responsible for selecting the system, determining the threshold, examining file integrity, interpreting the output, and approving the report. Documentation, record-keeping, audits, datasheets, and information concerning system use allow the process to be traced from the evidence and preprocessing stage to inference and interpretation [24], [31].

The assessment of deepfake detector outputs must integrate technical validity with the legality of their use. An accurate system may remain unsuitable when data processing violates the law, the method cannot be tested, or its use produces procedural unfairness. Administrative compliance cannot replace scientific validation [32]. The application of a reliability standard should assess the suitability of the method for the relevant evidence, the rate and distribution of errors, process traceability, the legality of data processing, openness to independent examination, and consistency with other evidence. The resulting assessment determines whether a detector output may be admitted, assigned limited evidentiary weight, or excluded from criminal proceedings.

3.2 Application of an AI Forensic Reliability Standard to Ensure Authenticity, Scientific Reliability, and Procedural Fairness

An AI forensic reliability standard is necessary to determine the admissibility and evidentiary weight of deepfake detector outputs. The assessment cannot depend on a

single score because each type of deficiency produces different legal consequences. Defects affecting the authenticity or legality of the evidentiary object may result in exclusion, while limitations concerning calibration, generalisation, or documentation primarily affect the weight of the expert opinion. The standard should encompass the authenticity of the evidentiary object, system suitability, data quality, error rates, robustness, traceability, expert competence, and procedural fairness.

Authenticity requires documentation of the source file, originating device or account, time of acquisition, imaging method, hash values, metadata, format alterations, and chain of custody. The analysis must be conducted on a forensic copy while preserving the original file. The correspondence of hash values after seizure establishes only the integrity of the copy, not the authenticity of the content before the file was obtained. Provenance and the context in which the content was produced must therefore remain subject to examination [33].

System suitability must be assessed by determining whether the detector is appropriate for the type of manipulation involved, including face swapping, lip synchronisation, facial reenactment, or voice cloning. The examination report must identify the system, version, configuration, reason for its selection, and operating conditions so that the result can be reproduced. Information concerning the training and validation data is also necessary to assess the representation of generators, demographic groups, resolutions, codecs, languages, and labelling procedures. Data that do not reflect the conditions of the evidentiary material may produce weak generalisation [9], [24].

Error rates should be presented through sensitivity, specificity, false-positive rates, false-negative rates, confusion matrices, decision thresholds, and uncertainty measurements under conditions resembling those of the case. A detector score must be explained as a classification output rather than as a probability of the defendant's guilt or error. Robustness must also be tested against compression, resizing, cropping, noise, transcoding, screen recording, and previously unseen generators because such alterations may change the resulting prediction [8], [10].

Traceability and procedural fairness require deepfake examinations to be documented, reproducible, and open to independent review. Experts must be competent, laboratories must apply quality control, and both parties must have access to the materials, reports, and testing process. Closed systems should receive limited evidentiary weight when confidentiality prevents meaningful scrutiny [30], [34]

Table 1. AI Forensic Reliability Standard Matrix

Dimension	Minimum Assessment Question	Required Documentation or Testing	Legal Consequence
Authenticity and legality of the evidentiary object	Was the file lawfully obtained, preserved without material alteration, and reliably connected to its source?	Original file or forensic image, hash values, metadata, acquisition records, imaging logs, provenance records, and chain-of-custody documentation	A material defect in authenticity or lawful acquisition may result in exclusion under Article 235 paragraph (5) of the Criminal Procedure Code
System suitability	Is the detector technically suitable for the relevant media format and the suspected type of manipulation?	System name, version, configuration, operating requirements, intended use, validation scope, and reasons for selecting the system	An unsuitable system may render the result irrelevant, inadmissible, or entitled to little evidentiary weight

Data quality and representativeness	Do the training and validation data sufficiently represent the characteristics of the evidence examined?	Dataset documentation, datasheets, data sources, demographic composition, generator coverage, labelling procedures, and train-validation-test separation	Poor representativeness weakens generalisation and reduces the evidentiary weight of the result
Error rates and calibration	Are the risks of false positives and false negatives known under conditions comparable to those of the case?	Sensitivity, specificity, false-positive rate, false-negative rate, confusion matrix, calibration curve, threshold, and uncertainty interval	A score presented without error-rate and calibration information cannot support a categorical conclusion
Robustness and generalisation	Does the result remain stable after transformations commonly affecting digital evidence and when applied to previously unseen generators?	Compression, resizing, cropping, noise, transcoding, screen-recording, adversarial, and cross-dataset tests	Failure of robustness or generalisation limits the scope of the expert conclusion and reduces evidentiary weight
Traceability and reproducibility	Can each stage of the examination be audited and independently repeated?	Acquisition records, preprocessing workflow, parameters, source files, logs, software environment, model configuration, and score-aggregation method	An untraceable or irreproducible process may justify reduced weight or exclusion
Expert competence and quality assurance	Does the expert possess relevant competence in digital forensics and machine-learning validation, and was the examination subject to quality control?	Curriculum vitae, training records, certification, laboratory procedures, proficiency testing, peer review, and quality-assurance records	Insufficient competence or quality assurance undermines the reliability and weight of the expert opinion
Independent and adversarial testing	Did the opposing party receive sufficient access and opportunity to verify, replicate, or challenge the examination?	Identical forensic copy, complete report, relevant logs, methodological information, comparison methods, and access for an independent expert	Denial of meaningful access may violate due process and justify disclosure orders, adjournment, reduced weight, or exclusion
Transparency and explainability	Are the basis, limitations, and interpretation of the detector output sufficiently understandable to the court and the parties?	Technical explanation, validation report, model documentation, interpretability outputs, limitation statement, and explanation of alternative causes	Inadequate transparency may prevent meaningful scrutiny and reduce the probative value of the result
Data protection and security	Was the processing of facial, vocal, or biometric data lawful, secure, necessary, and limited to the evidentiary purpose?	Lawful processing basis, access controls, storage records, security safeguards, retention policy, and vendor-processing information	Unlawful or insecure processing may affect admissibility, trigger data-protection liability, or require exclusion of improperly obtained material

Conflict of interest and vendor involvement	Were financial, institutional, or professional interests affecting the detector or expert fully disclosed?	Conflict-of-interest statement, vendor relationship records, licensing information, and disclosure of technical assistance	Undisclosed conflicts may reduce credibility and evidentiary weight
Consistency with other evidence	Is the detector output consistent with metadata, device artefacts, witness testimony, communications, and other forensic findings?	Comparative forensic analysis, metadata examination, device records, communication logs, witness evidence, and alternative detector results	An uncorroborated detector output cannot independently establish manipulation, attribution, intent, or criminal responsibility
Proportionality of the conclusion	Does the language of the expert conclusion remain within the limits supported by the data and method?	Written opinion, uncertainty statement, alternative explanations, and explanation of the proposition actually supported	Overstated or categorical conclusions may be restricted, assigned limited weight, or disregarded by the court
Accountability and responsibility	Can the persons responsible for selecting, operating, interpreting, and approving the system output be identified?	Audit trail, operator records, approval records, decision logs, and allocation of professional responsibility	Unclear responsibility weakens accountability and may reduce the reliability of the examination

Deepfake detectors should function primarily as investigative triage tools until their outputs are independently verified through forensic examination. Investigators must preserve the originating device or account, create forensic copies, maintain metadata, record hash values, and document the chain of custody to ensure that the examined file remains identifiable and traceable [35]. Prosecutors must subsequently assess the legality of acquisition, system suitability, error rates, operating conditions, methodological transparency, and any conflicting test results. Algorithmic labels cannot independently justify arrest, search, additional seizure, rejection of an alibi, or attribution to a particular person because a detector only assesses indications of manipulation. Identification of the perpetrator still requires supporting evidence such as device artefacts, account records, application logs, communications, transactions, and other materials demonstrating control, knowledge, and intent.

At trial, the original media, analytical visualisations, detector outputs, and expert opinions must be clearly distinguished to prevent technical demonstrations from creating an exaggerated impression of certainty [6]. Experts should explain their qualifications, the object examined, the system and version used, preprocessing procedures, validation data, performance metrics, decision thresholds, robustness, limitations, and possible alternative explanations. Courts may order independent testing, require additional disclosure, restrict the propositions that may be inferred, reduce evidentiary weight, or exclude results that cannot be meaningfully audited. Proprietary systems may still be used when sufficient methodological information, audit records, and access for independent defence experts are available, even if complete publication of the source code is unnecessary [31]. The applicable reliability standard should become stricter when detector outputs are used to reject an alibi, support coercive measures, or connect the defendant to disputed content. Indonesian forensic institutions should therefore develop validation datasets reflecting local devices, codecs, communication platforms, languages, and demographic characteristics, supported by periodically updated technical guidance [1], [6], [18], [23].

The legal consequence should correspond to the nature and seriousness of the deficiency. Unlawful acquisition or material defects affecting authenticity may result in exclusion under Article 235 paragraph (5) of the Criminal Procedure Code, while system incompatibility, insufficient validation, weak generalisation, or incomplete documentation generally reduce evidentiary weight or require supplementary examination [13]. A denial of meaningful access to files, logs, validation reports, or testing materials may also raise due process concerns and justify disclosure orders, adjournment, rebuttal examination, reduced weight, or exclusion. Courts must consistently distinguish manipulation classification from perpetrator attribution because even a reliable detector may indicate that media have been altered without proving the identity of the creator, the device used, intent, or purpose. Claims that authentic recordings are deepfakes should therefore be tested through provenance, metadata, witness testimony, device analysis, and validated detection methods without lowering the prosecution's burden of proving guilt through lawful and convincing evidence [3].

4. CONCLUSION

Based on the analysis of the legal status and reliability of deepfake detectors within the Indonesian criminal evidence system, their use must be carefully regulated to support truth-finding without impairing the rights of the parties. Deepfake detector outputs may be presented through electronic evidence, examination reports, and expert testimony, but they cannot function as independent evidence or as conclusive determinants of truth. Their probative value depends on the authenticity of the evidentiary object, system validity, data quality, error rates, model robustness, process traceability, expert competence, and the parties' opportunity to challenge the examination. Violations affecting legality or authenticity may result in the exclusion of evidence, while methodological weaknesses reduce the weight of expert testimony. The Supreme Court, police, prosecution service, forensic laboratories, system providers, and research institutions should establish guidelines concerning examination procedures, file preservation, local validation, version recording, independent testing, audit logs, and benchmarks adapted to the characteristics of digital media in Indonesia.

REFERENCES

[1] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Comput. Surv.*, vol. 54, no. 1, pp. 1–41, 2021, doi: 10.1145/3425780.

[2] R. Tolosana, S. Romero-Tapiador, J. Fierrez, and R. Vera-Rodriguez, "Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection," *Inf. Fusion*, vol. 64, no. 1, pp. 131–148, 2020, doi: 10.1016/j.inffus.2020.06.014.

[3] R. Chesney and D. K. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," *Calif. Law Rev.*, vol. 107, no. 6, pp. 1753–1820, 2019, doi: Chesney, Robert and Citron, Danielle Keats, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security (July 14, 2018). 107 California Law Review 1753 (2019), U of Texas Law, Public Law Research Paper No. 692, U of Maryland Legal Studies Research Paper No. 2018-21, Available at SSRN: https://ssrn.com/abstract=3213954 or http://dx.doi.org/10.2139/ssrn.3213954.

[4] M.-P. Sandoval, M. de Almeida Vau, J. Solaas, and L. Rodrigues, "Threat of deepfakes to the criminal justice system: A systematic review," *Crime Sci.*, vol. 13, 2024, doi: 10.1186/s40163-024-00239-1.

[5] D. Fallis, "The epistemic threat of deepfakes," *Philos. Technol.*, vol. 34, no. 4, pp. 623–643, 2021, doi: 10.1007/s13347-020-00419-2.

[6] J. T. Hancock and J. N. Bailenson, "The social impact of deepfakes," *Cyberpsychology, Behav. Soc. Netw.*, vol. 24, no. 3, pp. 149–152, 2021, doi: 10.1089/cyber.2021.29208.jth.

[7] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 5, pp. 910–932, 2020, doi: 10.1109/JSTSP.2020.3002101.

[8] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE Access*, vol. 10, pp. 25494–25513, 2022, doi: 10.1109/ACCESS.2022.3154404.

[9] T. Wang, X. Liao, K. P. Chow, X. Lin, and Y. Wang, "Deepfake detection: A comprehensive survey from

the reliability perspective," *ACM Comput. Surv.*, vol. 57, no. 3, pp. 1–35, 2025, doi: 10.1145/3699710.
[10] L. Stroebel, M. Llewellyn, T. Hartley, T. S. Ip, and M. Ahmed, "A systematic literature review on the effectiveness of deepfake detection techniques," *J. Cyber Secur. Technol.*, vol. 7, no. 2, pp. 83–113, 2023, doi: 10.1080/23742917.2023.2192888.
[11] A. L. Roth, "Machine testimony," *Yale Law J.*, vol. 126, no. 7, pp. 1972–2053, 2017.
[12] A. Biedermann and T. Lau, "Decisionalizing the problem of reliance on expert and machine evidence," *Law, Probab. Risk*, vol. 23, no. 1, 2024, doi: 10.1093/lpr/mgae007.
[13] "Undang-Undang Nomor 20 Tahun 2025 Tentang Kitab Undang-Undang Hukum Acara Pidana," Mar. 25, 2025. [Online]. Available: https://peraturan.go.id/id/uu-no-20-tahun-2025
[14] Republik Indonesia, *Undang-Undang Nomor 1 Tahun 2024 tentang Perubahan Kedua atas Undang-Undang Nomor 11 Tahun 2008 tentang Informasi dan Transaksi Elektronik*. 2024.
[15] S. A. Kusnadi and D. W. S. Putri, "Perlindungan hak privasi dalam penyalahgunaan teknologi deepfake di Indonesia," *J. Rechts Vinding Media Pembin. Huk. Nas.*, vol. 14, no. 2, pp. 195–210, 2025, doi: 10.33331/rechtsvinding.v14i2.2135.
[16] H. Antoro, D. Budhijanto, and Somawijaya, "Admissibility of artificial intelligence as electronic evidence: Comparative perspectives from Indonesia, the United States, and Japan," *J. IUS Kaji. Huk. dan Keadilan*, vol. 14, no. 1, pp. 122–139, 2026, doi: 10.29303/ius.v14i1.1880.
[17] B. van der Sloot and Y. Wagensveld, "Deepfakes: Regulatory Challenges for the Synthetic Society," *Comput. Law Secur. Rev.*, vol. 46, pp. 1–15, 2022, doi: 10.1016/j.clsr.2022.105716.
[18] E. Hydara, M. Kikuchi, and T. Ozono, "Empirical assessment of deepfake detection: Advancing judicial evidence verification through artificial intelligence," *IEEE Access*, vol. 12, pp. 151188–151203, 2024, doi: 10.1109/ACCESS.2024.3480320.
[19] P. M. Marzuki, *Penelitian Hukum*, Edisi revisi. Jakarta, Indonesia: Kencana, 2021.
[20] S. Soekanto and S. Mamudji, *Penelitian Hukum Normatif: Suatu Tinjauan Singkat*. Jakarta: Rajawali Pers, 2015.
[21] Republik Indonesia, *Undang-Undang Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi*. 2022.
[22] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, 2023, doi: 10.1016/j.patter.2023.100804.
[23] F. Abbas and A. Taeihagh, "Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence," *Expert Syst. Appl.*, vol. 252, 2024, doi: 10.1016/j.eswa.2024.124260.
[24] T. Gebru *et al.*, "Datasheets for datasets," *Commun. ACM*, vol. 64, no. 12, pp. 86–92, 2021, doi: 10.1145/3458723.
[25] M. Groh, Z. Epstein, C. Firestone, and R. Picard, "Deepfake detection by human crowds, machines, and machine-informed crowds," *Proc. Natl. Acad. Sci.*, vol. 119, no. 1, 2022, doi: 10.1073/pnas.2110013119.
[26] E. O. S. Hiariej, *Teori dan Hukum Pembuktian*. Jakarta, Indonesia: Erlangga, 2012.
[27] P. W. Grimm, M. R. Grossman, and G. V Cormack, "Artificial intelligence as evidence," *Northwest. J. Technol. Intellect. Prop.*, vol. 19, no. 1, pp. 9–106, 2021.
[28] R. Wexler, "Life, liberty, and trade secrets: Intellectual property in the criminal justice system," *Stanford Law Rev.*, vol. 70, no. 5, pp. 1343–1429, 2018.
[29] R. V Segate, "Cognitive bias, privacy rights, and digital evidence in international criminal proceedings: Demystifying the double-edged AI revolution," *Int. Crim. Law Rev.*, vol. 21, no. 2, pp. 242–279, 2021, doi: 10.1163/15718123-bja10048.
[30] K. M. Richmond, S. M. Muddamsetty, T. Gammeltoft-Hansen, H. P. Olsen, and T. B. Moeslund, "Explainable AI and law: An evidential survey," *Digit. Soc.*, vol. 3, 2024, doi: 10.1007/s44206-023-00081-z.
[31] J. A. Kroll *et al.*, "Accountable algorithms," *Univ. PA. Law Rev.*, vol. 165, no. 3, pp. 633–705, 2017.
[32] M. Kattnig, A. Angerschmid, T. Reichel, and R. Kern, "Assessing trustworthy AI: Technical and legal perspectives of fairness in AI," *Comput. Law Secur. Rev.*, vol. 55, 2024, doi: 10.1016/j.clsr.2024.106053.
[33] R. B. da Silva, "Updating the authentication of digital evidence in the International Criminal Court," *Int. Crim. Law Rev.*, vol. 22, no. 5–6, pp. 941–964, 2022, doi: 10.1163/15718123-bja10083.
[34] F. Romero-Moreno, "Deepfake Detection in Generative AI: A Legal Framework Proposal to Protect Human Rights," *Comput. Law Secur. Rev.*, vol. 58, p. 106162, 2025, doi: 10.1016/j.clsr.2025.106162.
[35] L. Freeman, "Weapons of war, tools of justice: Using artificial intelligence to investigate international crimes," *J. Int. Crim. Justice*, vol. 19, no. 1, pp. 35–53, 2021, doi: 10.1093/jicj/mqab013.