基于区块链与数字水印的 Deepfake 溯源技术应用场景研究

韦彩虹　唐钰婷　潘瑶婷　陈澳迪　杨小梅*
广西职业师范学院
DOI:10.32629/pe.v3i6.18008

[摘　要] 随着生成式人工智能技术的飞速发展,Deepfake等深度伪造技术已对网络舆论安全、司法证据体系与社会信任基础构成了严峻挑战。传统多媒体内容认证与溯源机制在去中心化传播环境中显得力不从心,存在易篡改、难追踪等固有缺陷。本文旨在系统探讨区块链技术与数字水印技术融合应用于Deepfake内容溯源的可行性、架构设计及其多元化应用场景。通过构建“嵌入-存证-验证”协同框架,本研究提出了一种能够覆盖内容生成、传播与取证全链条的可信溯源方案。论文进一步结合新闻媒体、司法存证、公共事务、金融商业及文化遗产五大典型领域,分析了差异化技术配置路径与治理策略。研究发现,技术融合能有效提升溯源精度与证据可信度,但其广泛应用仍面临法律标准缺失、性能瓶颈及协同治理机制待完善等挑战。研究结论为构建适应数字时代的网络舆论风险防控体系提供了兼具技术创新与制度设计的综合治理思路。
[关键词] 区块链；数字水印；深度伪造；内容溯源；网络治理；可信存证
中图分类号：TS872　**文献标识码**：A

Research on Application Scenarios of Deepfake Traceability Technology Based on Blockchain and Digital Watermarking

Caihong Wei　Yuting Tang　Yaoting Pan　Aodi Chen　Xiaomei Yang*
Guangxi Vocational Normal University

[Abstract] With the rapid advancement of generative artificial intelligence technology, deepfake and other deepfake technologies have posed severe challenges to online public opinion security, judicial evidence systems, and the foundation of social trust. Traditional multimedia content authentication and traceability mechanisms struggle to cope with decentralized dissemination environments, exhibiting inherent flaws such as susceptibility to tampering and difficulty in tracking. This paper systematically explores the feasibility, architectural design, and diverse application scenarios of integrating blockchain technology with digital watermarking for deepfake content traceability. By establishing a "embedding-proofing-verification" collaborative framework, this study proposes a trusted traceability solution capable of covering the entire chain from content generation to dissemination and evidence collection. The paper further analyzes differentiated technical configuration paths and governance strategies across five typical domains: news media, judicial evidence preservation, public affairs, financial commerce, and cultural heritage. Findings reveal that technological integration can effectively enhance traceability accuracy and evidence credibility, yet widespread adoption still faces challenges such as the absence of legal standards, performance bottlenecks, and incomplete collaborative governance mechanisms. The research conclusions provide a comprehensive governance approach combining technological innovation and institutional design to construct a digital-era risk prevention system for online public opinion.
[Key words] blockchain; digital watermarking; Deepfake; content provenance; internet governance; trusted evidence preservation

1 绪论

生成式人工智能技术的迅猛发展,在催生创新应用的同时,也带来了以Deepfake为代表的深度伪造技术滥用问题。这类技术能够合成高度逼真且难以肉眼辨别的虚假音视频,正日益成为冲击舆论公信力、侵蚀司法证据体系与破坏社会信任基础的重要工具。虚假信息凭借社交网络的放大效应得以快速、广泛

传播,其辨识难度高、溯源成本大,对国家安全、政治生态与社会稳定构成了现实且紧迫的威胁[1]。因此,构建一种能够对多媒体内容进行可靠身份认定与传播追溯的技术治理体系,已成为网络信息内容管理领域的一项核心挑战。

当前主流的应对策略集中于开发基于深度学习的被动检测算法。然而,这种技术路径存在固有局限：其一,检测模型的发展常滞后于伪造技术的演进,陷入持续的“攻防竞赛”；其二,算法依赖大量计算资源与标注数据,且其判定结果本身缺乏具有法律效力的存证支撑;其三,传统基于中心化数据库的数字水印或元数据管理方案,易受单点故障与内部篡改风险的影响,在去中心化的传播环境中显得脆弱。因此,探索一种融合主动标识与分布式存证的新型溯源范式,旨在从源头上为数字内容赋予不可抵赖的身份凭证,并对其流转过程进行可信记录,具有显著的实践必要性。

2 区块链与数字水印的技术融合基础

2.1区块链的技术内核及其在溯源中的适配性

区块链技术的核心价值在于其通过密码学哈希、链式结构与分布式共识建立的去中心化信任[2]。其不可篡改性确保了任何上链的存证信息(如内容哈希、时间戳、操作记录)一旦被网络确认,便极难被单独修改或删除,这为追溯信息提供了坚实的“信任锚点”。时间戳服务能为多媒体内容的生成、首次发布及关键传播节点提供精确且可信的时间证明。智能合约的引入,使得溯源流程可以自动化执行,例如,当验证请求触发合约时,可自动完成水印提取、哈希比对与结果返回,提升效率并减少人为干预。跨链技术的发展,则为不同平台、不同机构建立的独立溯源链之间的信息互通提供了可能,有助于形成覆盖更广的协同治理网络。

2.2数字水印技术的演进与鲁棒性挑战

数字水印技术通过将特定标识信息不可感知地嵌入多媒体载体中,使其成为内容的“数字指纹”。鲁棒水印旨在抵抗常见的信号处理操作(如压缩、滤波、尺寸调整),确保标识在内容流转过程中不丢失,适用于版权追踪。脆弱水印和半脆弱水印则对篡改敏感,一旦内容被非法修改,水印便会破坏或产生特定响应,从而精确定位篡改区域,适用于内容完整性认证[3]。近年来,基于深度学习的水印技术展现出更强的自适应嵌入与提取能力,能在复杂攻击下保持更好的鲁棒性。技术的核心挑战之一始终是在不可感知性(视觉/听觉质量)与鲁棒性(抗攻击能力)之间取得最佳平衡。

2.3融合协同机制与双层架构设计

二者的融合并非简单叠加,而是构建一个优势互补的有机整体。我们提出一个双层协同溯源架构：在第一层(内容层),利用数字水印技术,在内容创作或首次发布时,即将含有创作者身份、作品唯一编号或生成设备特征等信息的标识,作为“内在基因”嵌入音视频数据中。该水印应具备强鲁棒性,以应对网络传播中的各种转码与处理。在第二层(存证层),将上述水印的特征信息(或其哈希值)、内容的元数据(如生成时间、地点、哈希值)、以及关键的传播日志(如转发节点、时间)等信息,打包成交易记录在区块链上。区块链在这里充当了全局、可信的“公证簿”和“关系图谱”。二者通过哈希函数紧密联动：内容水印信息与其在链上的存证记录通过哈希值相互锁定。验证时,先从待测媒体中提取水印,获取其身份信息；再查询区块链,比对链上记录的该身份对应的原始内容哈希及其他元数据,从而完成对内容来源、完整性与传播历史的核验。此外,可结合零知识证明等密码学方案,在验证水印有效性的同时,不泄露水印的具体内容或创作者隐私,实现隐私保护下的可信溯源。

3 深度伪造溯源技术的应用场景建模

3.1新闻媒体与社交平台内容认证

场景特征表现为用户生成内容海量、传播速度快、实时性要求高。技术方案上,可设计轻量级水印算法,在用户上传视频时由平台客户端或服务器快速嵌入包含用户ID与时间戳的水印,并将水印哈希同步至一条由媒体平台、权威机构等共同维护的联盟链上[4]。面向普通用户,可开发浏览器插件或移动端APP,使其能一键对可疑视频进行初步水印探测与链上验证,结果以“可信”、“存证可疑”或“无溯源信息”等标签形式提示。治理上,平台可将溯源认证结果与内容推荐权重、流量分配机制挂钩,激励原创和真实内容传播,形成“技术验证-社区评价-平台调节”的协同治理生态。

3.2司法电子证据存证与溯源

司法场景对证据的完整性、真实性与法律效力要求极高。技术方案需符合司法合规标准,可采用符合国家密码管理要求的数字水印算法,并在取证过程中即时将音视频证据的哈希值、水印信息、取证时间、取证人员及设备信息等固定至司法区块链存证平台。该平台应由法院、公证处、鉴定中心等权威节点共同运维,确保证据链的全过程可信。当法庭需要对涉案的Deepfake视频进行溯源时,可授权鉴定机构调取链上存证,与当庭出示的证据进行比对,快速认定证据是否被篡改及其原始来源,极大提升电子证据的采信效率与司法审判的公正性。

3.3政治选举与公共事务信息治理

此场景具有高风险、跨国界、强动机的特点,关乎国家安全与社会稳定。技术方案应强调权威性与国际协作。可由政府主导,联合主流媒体、社交平台及国际合作伙伴,建立国家级甚至国际性的“重要公共信息认证联盟链”。对于总统辩论、官方新闻发布会等关键音视频,在发布前即嵌入高安全等级的数字水印,并即时上链存证。一旦出现伪造内容在社交媒体传播,各参与方节点可快速响应,利用链上存证进行验证与辟谣。同时,可利用脆弱水印技术对国家档案馆的重要历史影像资料进行保护,任何篡改都会被系统识别并告警。

3.4金融与商业领域可信数据交换

金融商业场景对身份认证与合同安全的依赖极强。技术方案可聚焦于远程视频面签、商业合同谈判录像、机密商业演示等场景。结合生物特征水印技术,可将对话者的声纹特征或经过处理的生物模板嵌入会话录音或视频中,增强身份绑定。同时,

利用智能合约设定访问权限,只有授权方才能触发水印提取与验证流程,并将验证结果作为后续交易或合同执行的自动化条件。这为远程开户、线上信贷审批、跨境商业谈判等业务提供了兼具便利性与安全性的可信基础。

3.5文化遗产数字资源保护

博物馆、档案馆等机构拥有大量高价值的数字化文物、古籍和艺术品影像。技术方案旨在实现长期保真与授权管理。可为每一份数字藏品嵌入不可见的版权与所有权水印,并将其唯一标识和元数据存证于区块链。当藏品被授权用于出版、展览或衍生品开发时,授权记录(被授权方、期限、用途)可通过智能合约上链管理,实现授权流转的全程追溯。这不仅能有效防止数字文物被非法复制和滥用,也为数字文创产品的版权交易与收益分配提供了透明、高效的解决方案。

4 技术实施挑战与治理路径优化

4.1技术层面的挑战与应对

首先,技术性能存在平衡难题。强鲁棒水印往往需要较大的嵌入容量,可能在高压缩率(如短视频平台)下影响视觉质量或导致信息丢失。需要研发更适应网络传播特性的自适应水印算法。其次,海量多媒体数据直接上链成本高昂。可行的解决方案是采用“链上-链下”混合存储,仅将内容哈希、水印密钥摘要等核心指纹信息上链,原始媒体文件存储在分布式文件系统(如IPFS)或受信云中,通过哈希值进行锚定。再次,区块链的交易确认延迟(从几秒到数分钟不等)与虚假信息的秒级传播速度存在矛盾。可在热点事件预警机制中,对关键信源实施“预存证”和快速验证通道[5]。

4.2法律、标准与协同治理瓶颈

技术方案的有效性最终依赖于法律环境的认可与标准化框架的支撑,而当前相关领域的法规与标准建设明显滞后。在司法层面,经由区块链与数字水印处理的电子证据,其法律效力认定标准尚未统一,取证操作流程的规范性与鉴定意见的采信规则有待明确,这影响了技术在诉讼中的广泛应用。在国际层面,深度伪造内容的跨境传播使得溯源工作必然涉及不同法域,面临数据主权法规冲突、司法管辖权争议以及执法协作机制不畅等系统性障碍。行业技术标准的缺失同样制约了规模化应用,包括水印算法的互操作性标准、区块链存证的数据格式规范、不同平台间验证结果的互认机制等均未建立,导致容易形成新的“技术孤岛”。破解这些困境,需要在国家层面加快推动专项立法与行业标准制定,明确技术应用的法律边界与程序要求;在国际层面,则需依托多边对话机制,探索建立基于对等原则的跨境数据取证与司法协作示范框架,为全球性治理提供制度参照。

5 总结

本研究系统论证了区块链与数字水印技术融合应用于Deepfake内容溯源的可行性与应用潜力。通过构建双层协同架构,该方案能够为多媒体内容赋予不可剥离的“数字身份”,并在去中心化的网络中记录其完整的“生命轨迹”,从而为识别和打击深度伪造提供了从技术到存证的一体化解决方案。针对不同应用场景的差异化建模表明,不存在“一刀切”的万能方案,需根据具体场景的风险属性、性能要求和参与主体,灵活配置技术参数与治理规则。研究的最终有效性,不仅取决于技术本身的持续进化,更依赖于法律法规的配套、技术标准的建立以及社会多元主体的协同共治。

[项目信息]

本项目由国家级大学生创新创业训练计划项目资助,项目名称：广西职业师范学院2025年大学生创业训练计划项目《生成式人工智能Deepfake技术的网络舆论风险防控与治理研究》,项目级别：国家级,项目类别：一般项目,项目编号：202514684005。

[参考文献]

[1]刘立伟,傅超豪,孙泽堃,等.数据要素流通全流程隐私关键技术:现状、挑战与展望[J/OL].软件学报,1-25[2025-12-07].https://doi.org/10.13328/j.cnki.jos.007478.

[2]卫霞,白国柱,张文俊.基于区块链技术对抗深度伪造现状研究[J].信息安全研究,2021,7(07):615-620.

[3]陈泊睿,张梅,李昕蕊,等.数字水印技术原理与发展[J].中国防伪报道,2025,(10):98-102.

[4]孙闻阳.新闻溯源:人工智能内容标识的策略[J].新闻文化建设,2025,(03):45-47.

[5]陆晨旭.基于指纹特征的深度伪造人脸检测与溯源研究[D].杭州电子科技大学,2025.

作者简介：

韦彩虹(2002--),女,壮族,广西都安瑶族自治县人,本科,单位：广西职业师范学院,研究方向：人工智能。

唐钰婷(2003--),女,汉族,广西容县人,本科,单位：广西职业师范学院,研究方向：人工智能。

潘瑶婷(2004--),女,壮族,广西贵港人,本科,单位：广西职业师范学院,研究方向：人工智能。

陈澳迪(2002--),女,汉族,广西钦州人,本科,单位：广西职业师范学院,研究方向：人工智能。

*通讯作者：

杨小梅(1981--),女,汉族,山东青岛人,博士研究生,单位：广西职业师范学院,研究方向：计算语言学、人工智能。