

XAI-IPFD: An Evidence-Aware Explainable Artificial Intelligence Architecture For Intellectual Property Fraud And Digital Piracy In The Orange Economy

Atiku Baba Shidawa*¹, Abednego Gambo Habu¹, Achi Aaron Unimke²,
Benjamin Opabunmi¹, Krulat Naandi Dariyem¹, Theresa Awulor³

¹National Institute for Policy and Strategic Studies, (NIPSS), Nigeria

²Higher School of Economics, Moscow.

³National Open University of Nigeria, Nigeria

ABSTRACT

Digital platforms have expanded creative markets while also increasing the speed and scale of unauthorised copying, deceptive ownership claims and cross-platform redistribution. Artificial intelligence can assist by identifying suspicious similarities, provenance anomalies and distribution patterns, but a detection score is not, by itself, an adequate basis for consequential intellectual property decisions. This limitation is especially important in the Orange Economy, where works are frequently multimodal, ownership may be shared, licences can be territorial, derivative use may be legitimate, and creators may themselves face fraudulent or mistaken rights claims. This article develops XAI-IPFD, a literature-grounded, model-agnostic and evidence-aware architecture for explainable intellectual property fraud and digital piracy decision support. An integrative synthesis of recent work on AI-based detection, watermarking, provenance technologies, generative-AI copyright risk and human-centred explainable AI identifies a persistent integration gap: prediction, provenance and explanation are commonly treated as separate technical functions, while rights context, evidence correlation, mandatory human verification and auditable case construction remain weakly integrated. XAI-IPFD responds through eight interoperable layers covering lawful evidence sources, preprocessing, feature representation, detection and uncertainty, evidence correlation, explainability, human verification and auditable output. The architecture treats LIME, SHAP and ANCHOR as complementary explanation families and formalises the case record, rather than the prediction alone, as the unit of accountability. It also incorporates territorial rights, multi-party attribution, derivative-use ambiguity and bidirectional protection against infringement and abusive claims. Conceptual validation combines design-requirement coverage, scenario walkthroughs and falsifiable research propositions without presenting internal coherence as empirical performance. The article contributes a sociotechnical account of explainable IP protection in which trustworthy decision support depends on linking machine signals to contextual evidence, intelligible explanations, authorised human judgement and reproducible audit records.

Keywords: explainable artificial intelligence; intellectual property fraud; digital piracy; Orange Economy; copyright; evidence-aware AI; human-in-the-loop; SHAP; LIME; ANCHOR; contestability.

INTRODUCTION

The infrastructure of creative production has changed markedly. Music, film, publishing, games, visual art and other creative products now move through digital platforms that allow creators to reach audiences with fewer physical and geographic constraints. The same infrastructure also makes copying, alteration, re-uploading and redistribution easier to scale. A disputed work can appear on several services and in

several jurisdictions before the creator, platform or rights administrator has enough information to determine what happened.

Artificial intelligence has become an important part of the technical response. Machine-learning systems can compare large content collections, identify suspicious websites, detect recurring patterns and assist with similarity analysis. Yet the institutional problem begins where the prediction ends. A high similarity

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

score does not reveal whether a valid licence exists, whether the claimant holds the relevant right, whether the use falls inside an authorised territory or whether a transformed work requires closer human interpretation. A system that moves directly from prediction to action therefore risks confusing technical suspicion with an evidential or legal conclusion.

This problem is increasingly visible in generative-AI environments. Copyright questions can arise during access to training material, model development, generation and downstream reuse, and the legal significance of those stages is not uniform across jurisdictions [7], [10], [15]. At the same time, explainable AI research has moved away from the assumption that opening a model's black box is sufficient. Human-centred XAI emphasises the needs, expertise and decision context of the people who must interpret an explanation [12], [4].

Against this background, the article treats digital IP protection as an evidence-aware decision-support problem. XAI-IPFD does not introduce a new classifier and does not report experimental performance. Instead, it specifies an architecture that connects prediction with uncertainty, external evidence, rights context, explanation, human verification and auditability. The contribution is therefore architectural and sociotechnical. It asks what a responsible IP decision system must preserve between the moment an AI model flags a case and the moment an authorised person decides what operational action, if any, is justified.

A. RESEARCH PROBLEM AND GAP

The literature on digital IP protection is technically diverse but functionally fragmented. Detection systems identify suspicious content or distribution behaviour. Watermarking provides traceable identifiers. Blockchain and other provenance mechanisms strengthen registration, timestamping

and transaction records. XAI methods interpret complex model behaviour. These streams solve different parts of the same institutional problem, yet they are not naturally interchangeable. Similarity is not infringement, provenance is not detection, and explanation is not evidence.

The gap addressed by XAI-IPFD is therefore an integration gap rather than an absence-of-research claim. Existing work demonstrates that explainable detection, provenance and content protection are individually feasible. What remains less developed is an end-to-end architecture in which a prediction is systematically connected to a structured evidence package, multi-party and territorial rights context, audience-appropriate explanations, non-bypassable human verification and a reproducible audit record. The architecture also addresses a second asymmetry: many protection systems are designed to assist the claimant, while creators may also need protection against mistaken or fraudulent claims. XAI-IPFD consequently treats bidirectional protection as a core requirement.

B. AIM AND RESEARCH QUESTIONS

The study aims to develop a literature-grounded architecture for explainable detection and review of intellectual property fraud and digital piracy in the Orange Economy. It addresses four questions. RQ1 asks which technical approaches are currently used for digital IP protection and piracy detection. RQ2 examines the limitations that prevent those approaches from supporting transparent, evidence-aware and contestable decisions. RQ3 considers which explanation mechanisms are suitable for technical and non-technical reviewers. RQ4 asks how detection, evidence correlation, explainability and human verification can be integrated into a reusable decision architecture.

C. CONTRIBUTIONS

Contribution	Specific contribution
Integrative synthesis	Connects detection, watermarking, provenance, generative-AI copyright risk and human-centred XAI as components of one decision problem.

Architecture	Defines an eight-layer, model-agnostic path from lawful acquisition to a human-reviewed and auditable case record.
Formalisation	Defines the case record, not the prediction alone, as the unit of accountability.
Domain specificity	Builds multimodality, attribution chains, territorial rights, derivative works and fraudulent claims into the design.
Governance	Makes contestability, proportionality, human verification and audibility functional system requirements.
Research programme	Links propositions to layers and measurable outcomes for subsequent technical and human-centred validation.

II. Evidence Base and Synthesis Method

The article uses a structured integrative synthesis rather than claiming a completed systematic review. The original evidence base was assembled primarily through a documented IEEE Xplore search and supplemented with directly relevant peer-reviewed journal, conference and legal-policy literature. For this submission version, the synthesis is strengthened with recent human-centred XAI research and contemporary copyright scholarship. The methodological boundary remains explicit: the purpose is theory-informed architecture design, not an exhaustive estimate of the global literature.

The synthesis follows four stages. First, studies are grouped by the function they perform in an IP-protection workflow. Second, each stream is examined for its input, technical output, evidential role, explanation capability and human-review assumptions. Third, cross-cutting limitations are identified, particularly where multimodality, rights context, traceability and contestability are weak. Fourth, those limitations are translated into design requirements. This functional approach is consistent with human-centred XAI scholarship that treats explanation as a sociotechnical design problem rather than a purely algorithmic property [12], [4].

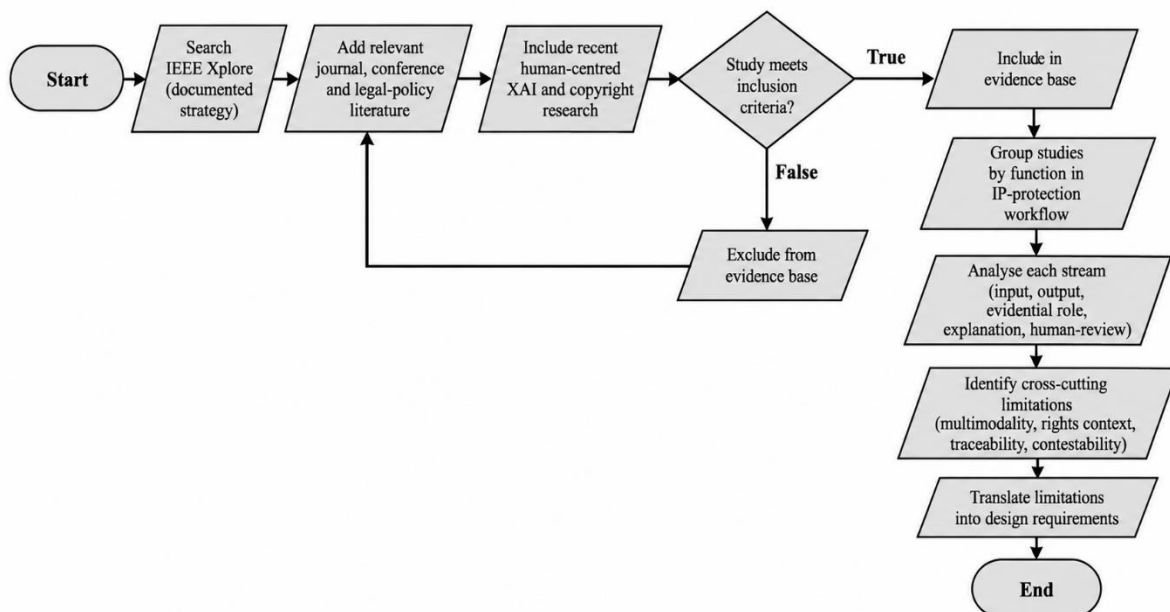


Figure 1. Evidence Base and Structured Integrative Synthesis Methodology for the Development of the XAI-IPFD Framework

Review element	Treatment in this article
Purpose	Identify capabilities, limitations and design implications for explainable IP decision support.
Temporal emphasis	Recent research concentrated on 2021-2026, with foundational sources retained when conceptually necessary.
Evidence streams	Detection; watermarking; provenance/rights management; generative-AI copyright; XAI; human-centred evaluation.
Synthesis unit	Technical function, evidential role, explanation role, human-review requirement and domain limitation.
Claim boundary	Patterns are attributed to the assembled literature; exhaustive global coverage is not claimed.
Future extension	A PRISMA-based multi-database review is specified as the first stage of subsequent validation.

III. State of the Art and Unresolved Integration Problem

A. AI-Based Detection and Similarity Analysis

AI-based protection systems can classify piracy-associated websites, identify suspicious textual or visual similarity and process volumes that would be difficult to inspect manually. Deep-learning approaches have also been applied to copyright recognition in specific media settings, including comics and visual content [3], [6]. Their principal limitation for the present problem is not necessarily predictive capability. It is that a bounded detector normally answers a bounded technical question. It can estimate whether content resembles a reference item, but it cannot independently establish the scope of a licence, the credibility of a claimant or the legal significance of a transformation.

This limitation makes uncertainty especially important. A practical detector should not return only a class label. It should expose a calibrated score or uncertainty measure that allows an institution to distinguish routine non-escalation from ambiguous cases requiring closer review. An item below a configured threshold should therefore be described as not automatically escalated, not as proven legitimate.

B. Watermarking and Forensic Traceability

Watermarking solves a separate aspect of the problem in that it embeds data which will be useful later in source identification, authentication, or ownership verification. Recent literature shows how effective methods have become in deep learning watermarking and forensic tracking [9], [8], [16]. While useful when the content has been deliberately watermarked beforehand, this cannot be taken for granted when dealing with legacy material, user-created content, or artistic creations that were never through such process. XAI-IPFD sees watermark data as evidence when available and not as a universal requirement.

C. Blockchain, Provenance and Rights Records

Blockchain technologies could serve to enhance provenance through registration, timestamps, license information and tamperproof history [1], [5]. These solutions do not provide diagnostics, however; all they offer is evidentiary value in that a blockchain will confirm the existence of a record on a certain date, but not necessarily that circulating material is infringed and/or that the owner recorded owns all necessary rights.

D. Generative AI and Copyright Complexity

Generative AI broadens the scope of copyright issues since conflicts may arise during training, license-

building, and results generation phases. Modern literature examines legal access, exemptions for text-and-data mining, transparency obligations and copyright infringement issues throughout the development process [7], [10], [15]. From an architectural point of view, a binary classification of 'infringing' is inadequate. A convincing record of the case must retain chronology, transformation, source, scope of any license, and the identity and qualifications of the rights holder making the claim.

E. Human-Centered Explainability

XAI solves model opacity problems, yet a technically accurate explanation may fail if it is not intelligible

and actionable to its audience. According to [12], human-centered XAI is a sociotechnical approach that focuses on user and contextual considerations, reflection and actionability. Recent systematic research reveals that the design of user-centered XAI is currently scattered and that explanations requirements differ depending on stakeholder role and expertise [4], [17]. Both studies imply adopting a stratified explanation solution instead of one interface for explanation in developers', rights holders', investigators' and authors' interests.

F. Comparative State of the Art

Research stream	Primary capability	Typical strength	Persistent limitation for XAI-IPFD
AI detection	Similarity/classification	Scalable automated screening	Prediction may lack rights context and contestable evidence.
Watermarking	Embedded traceability	Strong source/authenticity signal when marked	Cannot cover all legacy or unmarked works.
Blockchain/provenance	Registration and record integrity	Timestamped, tamper-resistant records	Does not itself detect infringement or interpret licence scope.
Post-hoc XAI	Model interpretation	Makes influential features/rules visible	Explanation is not external evidence or legal proof.
Human-centred XAI	User/context-sensitive explanation	Improves alignment with reviewer needs	Requires domain-specific workflow integration.
XAI-IPFD	Evidence-aware decision architecture	Integrates prediction, rights context, explanation, review and audit	Requires empirical prototype and institutional validation.

IV. Explainability Strategy

XAI-IPFD treats LIME, SHAP and ANCHOR as complementary explanation families. LIME creates a

local surrogate around an individual prediction and can support intuitive visual or instance-level inspection, although perturbation choices can affect stability. SHAP uses Shapley-value principles to

attribute feature contributions and is useful for technical audit, but high-dimensional explanations can be computationally demanding and difficult to communicate. ANCHOR generates local conditional rules that can be easier for non-technical reviewers to read, although strong precision inside an anchor does not imply broad coverage.

No technique is treated as universally superior. The choice depends on modality, detector, stakeholder and decision task. This approach is consistent with human-centred XAI research, which increasingly emphasises audience, context and the distinction between technical explanation quality and human usefulness [12], [4].

Dimension	LIME	SHAP	ANCHOR	XAI-IPFD requirement
Form	Local surrogate/visual importance	Feature attribution	Local conditional rule	Allow complementary views
Strength	Accessible local inspection	Structured contribution analysis	Readable rule communication	Match explanation to reviewer
Risk	Perturbation sensitivity	Cost and dimensional complexity	Limited rule coverage	Record limitations and parameters
Evaluation	Local fidelity/stability	Fidelity/stability	Precision/coverage	Also test comprehension and appropriate reliance

V. Design Requirements for the Orange Economy

The Orange Economy introduces rights structures that generic content matching can easily compress. A film may contain screenplay, music, performance, images and licensed clips. A song may involve composition, recording, samples and remixes. Rights can be

divided among creators, producers, publishers, collecting societies, licensees and distributors. The same work may be authorised in one territory and restricted in another. The architecture therefore has to preserve both technical similarity and the context that determines what that similarity means.

ID	Requirement	Rationale
DR1	Multi-format support	Creative products span text, image, audio, video, software and mixed media.
DR2	Explainable outputs	Reviewers need intelligible reasons and uncertainty, not only scores.
DR3	Structured evidence correlation	Predictions must connect to metadata, provenance, rights and similarity evidence.
DR4	Mandatory human verification	Consequential operational action remains reviewable and contestable.

DR5	Separate prediction and explanation evaluation	Accurate detection can coexist with poor explanation and vice versa.
DR6	Modularity	Representations, detectors and XAI methods should be replaceable.
DR7	Decision-support boundary	The architecture supports investigation and moderation, not final legal adjudication.
DR8	Multi-party and territorial rights context	Attribution, licence scope, territory and timing can alter a match's meaning.
DR9	Bidirectional protection	Evidence should support legitimate claims and help challenge mistaken or fraudulent claims.

VI. XAI-IPFD Architecture

XAI-IPFD comprises eight interoperable layers. The sequence is designed for accountability rather than computational rigidity. Implementations may

parallelise some processing, but every final case must remain traceable to source authority, preprocessing, feature representation, model version, evidence sources, explanation method and human review.

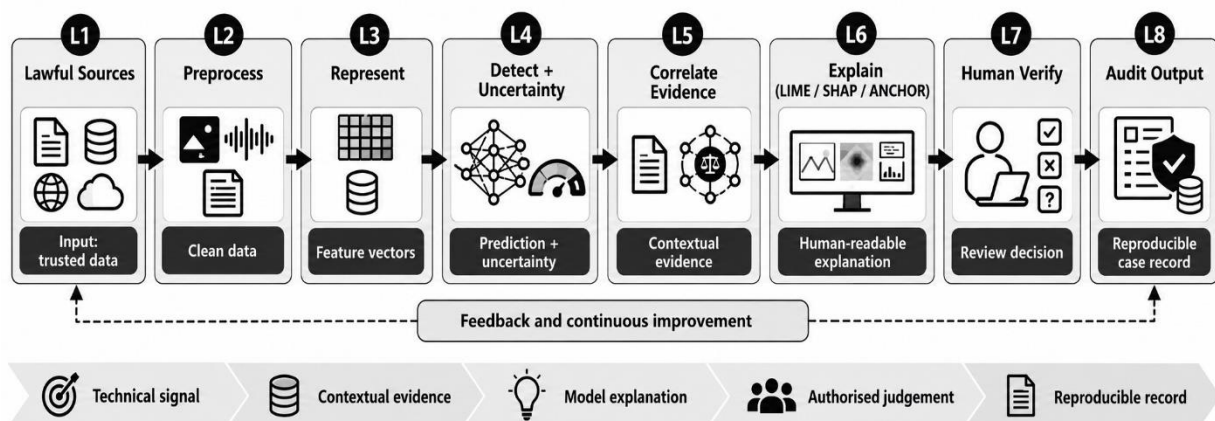


Figure 1. XAI-IPFD evidence-aware architecture.

A. Formal Case Representation

The architecture is formalised around a case object rather than a prediction alone. For case i , the conceptual record is $C_i = \{P_i, U_i, E_i, R_i, X_i, H_i, A_i\}$. This descriptive case object $C_i = \{P_i, U_i, E_i, R_i, X_i, H_i, A_i\}$, is converted into an active decision-logic model. we define a probabilistic Bayesian inference framework for calculating the infringement escalation risk score $R_{\text{case}}(i) \in [0,1]$. The layer parameters are defined as thus:

- **Detection Probability (P_i):** Model output score $P_i \in [0,1]$ indicating technical similarity.

- **Model Uncertainty (U_i):** Epistemic + aleatoric uncertainty measure $U_i \in [0,1]$ (where $U_i = 0$ implies absolute confidence)
- **Contextual Evidence Vector (E_i):** $E_i = \{e_{\text{provenance}}, e_{\text{watermark}}, e_{\text{chronology}}\} \in [0,1]^3$, where:

$e_{\text{provenance}}$: Validated timestamped registration record.

$e_{\text{watermark}}$: Verified forensic watermark signal.

$e_{\text{chronology}}$: Historical upload/prior publication timeline weight [1].

- **Rights & Licensing Weight (R_i):** $R_i \in [-1,1]$, representing legal authorization:

$R_i = +1$: Proven lack of authorization / invalid territory / expired license [1].

$R_i = 0$: Unknown / ambiguous licensing rights [1].

$R_i = -1$: Explicit valid license for territory, period, and media type [1].

The mathematical formulation of risk score, given as the combined probability of required operational escalation,

$R_{\text{case}}(i)$, is defined as:

$$R_{\text{case}}(i) = \sigma(w_p \cdot \hat{P}_i + w_e \cdot \bar{E}_i + w_r \cdot R_i)$$

Where:

- $\hat{P}_i$ (**Uncertainty-Adjusted Prediction**):

$$\hat{P}_i = P_i \cdot (1 - U_i) + 0.5 \cdot U_i$$

$$\text{Action}(C_i) = \begin{cases} \text{Auto-Dismiss / Log,} & \text{if } R_{\text{case}}(i) < \tau_{\text{low}} \\ \text{Mandatory Human Review (L7),} & \text{if } \tau_{\text{low}} \leq R_{\text{case}}(i) \leq \tau_{\text{high}} \\ \text{High-Priority Review \& Pre-Flag,} & \text{if } R_{\text{case}}(i) > \tau_{\text{high}} \end{cases}$$

Where τ_{low} and τ_{high} are empirically calibrated thresholds based on the cost-of-error matrix for false positives (abusive claims) versus false negatives

(As model uncertainty $U_i \rightarrow 1$, the prediction score reverts toward neutral risk 0.5, preventing overconfident automated decisions) [1]

- $\bar{E}_i$ (Composite Evidence Score):

$$\bar{E}_i = 1 - (\alpha \cdot e_{\text{provenance}} + \beta \cdot e_{\text{watermark}} + \gamma \cdot e_{\text{chronology}}), \quad \text{with } \alpha + \beta + \gamma = 1$$

- $\sigma(z)$ (Sigmoidal Logistic Function):

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

- w_p, w_e, w_r are non-negative weights calibrated such that $w_p + w_e + w_r = 1$.

Decision boundary and layer 7 escalation logic

The calculated $R_{\text{case}}(i)$ dictates the automated routing within Layer 7 (Human Verification) [1]:

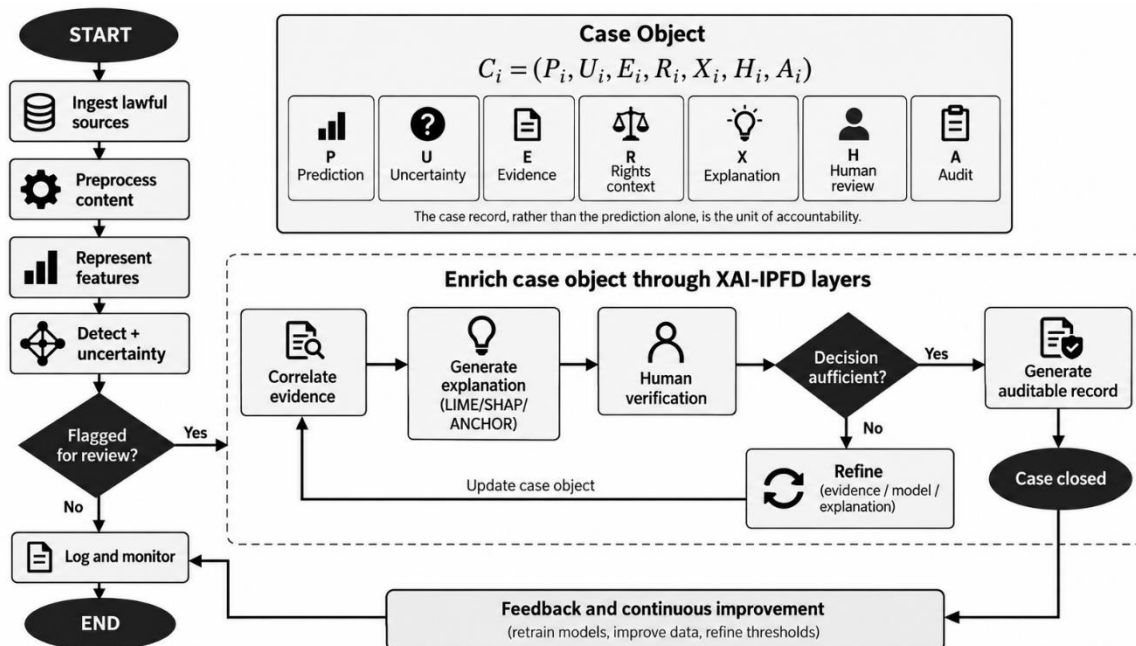


Figure 2. Formal case object used as the unit of accountability.

B. L1-L3: Lawful Inputs, Preprocessing and Representation

L1 identifies the content under examination and the authority under which supporting information is accessed. Candidate sources include authorised reference works, platform metadata, registration records, upload histories, licences, watermark indicators and chain-of-custody information. L2 standardises heterogeneous inputs while preserving provenance. Audio may be resampled, images resized, text tokenised and video separated into visual and audio streams, but quality-control decisions and exclusions remain logged. L3 creates modality-appropriate representations, including fingerprints and spectral descriptors for audio, perceptual hashes and learned embeddings for images, semantic representations for text, and temporal or multimodal representations for video.

Metadata remains distinguishable from content features. This separation matters because a reviewer should be able to tell whether a result was driven by characteristics of the work itself or by contextual signals such as upload history, distribution behaviour or rights metadata.

C. L4: Detection and Uncertainty

L4 is model-agnostic. CNNs, transformers, similarity networks, ensembles or specialised detectors can be selected according to modality and task. The architectural requirements are reproducible model identity, documented configuration, calibrated uncertainty and an explicit threshold policy. A practical deployment may use high-confidence, low-confidence and intermediate review bands, but thresholds must be empirically calibrated for the selected dataset and operational cost of error.

The language attached to these bands is deliberately cautious. Falling below an escalation threshold means that no automated escalation is generated; it does not establish legitimacy. Likewise, a high score establishes technical suspicion, not infringement.

D. L5: Evidence Correlation and Rights Context

L5 transforms an isolated prediction into a contextualised case by connecting content similarity to metadata, provenance, claimant identity, licence

scope, territory, chronology and other relevant evidence. The aim is not to maximise the number of evidence items but to preserve their meaning and provenance. A high similarity score may be expected where a valid licence covers the disputed distribution. A moderate score may become more significant when combined with earlier registration, a matching watermark and a suspicious upload sequence. In a fraudulent-claim scenario, chronology may weaken rather than strengthen the claimant's position.

This layer is also where the Orange Economy becomes more than a label. Composite works can contain separately owned components, and rights may differ across media, territories and periods. L5 therefore permits the case object to represent multiple rights holders and licence conditions instead of assuming a single claimant and a globally uniform entitlement.

E. L6: Explainability

L6 explains the L4 prediction rather than the legal status of the work. Each explanation report should identify the method, relevant parameters, influential features or rules, direction of contribution and known limitations. Multiple explanation forms may be used where they answer different reviewer needs, but agreement among LIME, SHAP and ANCHOR is not treated as independent evidence because each remains an interpretation of the same underlying model behaviour.

F. L7: Mandatory Human Verification

L7 places an authorised reviewer between automated analysis and consequential operational action. The reviewer receives prediction, uncertainty, evidence and explanation together and may support escalation, challenge the automated result, request additional evidence or dismiss the case. The rationale is recorded. Human review is not presumed infallible: expertise, workload, incentives and automation bias can influence decisions. For that reason, reviewer role, decision and rationale become part of the auditable case object.

G. L8: Auditable and Contestable Output

L8 preserves the complete decision-support record: content identifier, source authority, preprocessing

configuration, model version, detection score and uncertainty, evidence package, explanation output, reviewer decision, rationale and timestamps. A

tamper-resistant registry can be added where useful, but blockchain is optional. Auditability depends first on completeness, reproducibility and access control.

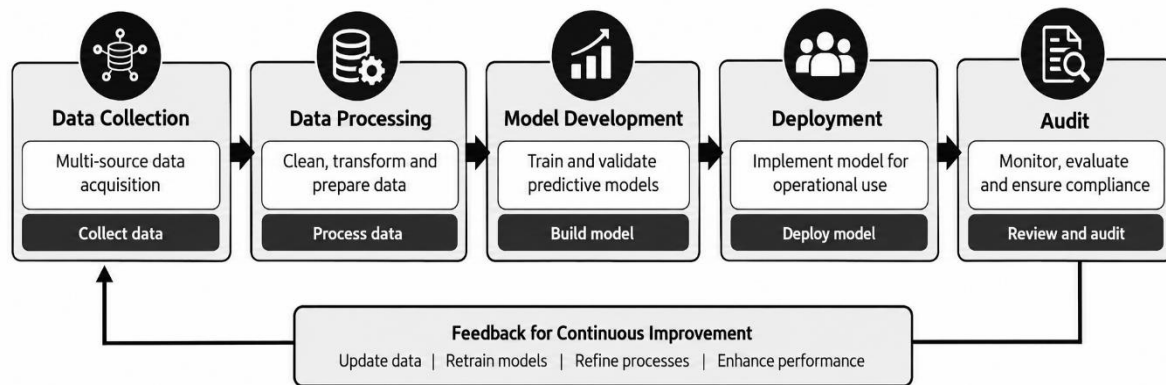


Figure 3. Separation of prediction, evidence, explanation, judgement and audit.

VII. Security, Governance and Responsible Deployment

Operational deployment would require controls for adversarial manipulation, data poisoning, model drift, privacy, access management and evidential integrity. Creative-content detectors are attractive targets for evasion because content can be deliberately transformed to remain recognisable to people while reducing machine similarity. Model maintenance should therefore be treated as a governed lifecycle rather than a one-time deployment task.

Explainability also creates a security trade-off. Contestability requires enough disclosure for a reviewer or affected party to understand the basis of a decision, while excessive technical detail may reveal features that facilitate evasion. The appropriate design

response is proportionate and role-based disclosure, not blanket opacity. Recent human-centred XAI research similarly emphasises that explanation should be designed around stakeholder roles and expertise rather than delivered identically to every user [17], [12].

The architecture maintains a legal boundary throughout. It supports investigation, moderation and institutional review but does not determine liability, fair use, admissibility or the final scope of a copyright entitlement. Contemporary generative-AI scholarship reinforces the need for this boundary because copyright obligations vary across legal systems and stages of the AI lifecycle [10], [15].

VIII. Scenario Walkthroughs

Scenario	Machine signal	Contextual evidence	Human-review question
Streaming audio piracy	Fingerprint/spectral similarity	Release, licence, distribution history, watermark	Is the match unauthorised or explained by a valid distribution right?
Modified visual work	Perceptual/deep similarity	Prior publication, registration, transformation history	Does the transformation require escalation or closer rights analysis?
Fraudulent rights claim	Contested similarity	Claimant identity, chronology, prior authorship and licence	Does the claimant's evidence support the asserted entitlement?

Territorial licence dispute	Content match	Territory, licence period, distribution location	Was the use authorised in the relevant place and period?
-----------------------------	---------------	--	--

These scenarios are conceptual walkthroughs rather than experiments. Their purpose is to test representational adequacy: whether one architecture can preserve the information required for technically different disputes without changing its core structure. They do not establish predictive accuracy, legal admissibility or institutional effectiveness.

IX. Conceptual Validation

The present article validates XAI-IPFD at the level appropriate to a conceptual architecture. Three forms

of evidence are combined. First, requirement coverage asks whether each problem identified in the synthesis has an explicit architectural response. Second, scenario walkthroughs test whether the architecture can represent materially different IP disputes. Third, falsifiable propositions translate the architecture into claims that can later be tested experimentally. This layered approach strengthens conceptual validity without presenting internal coherence as empirical performance.

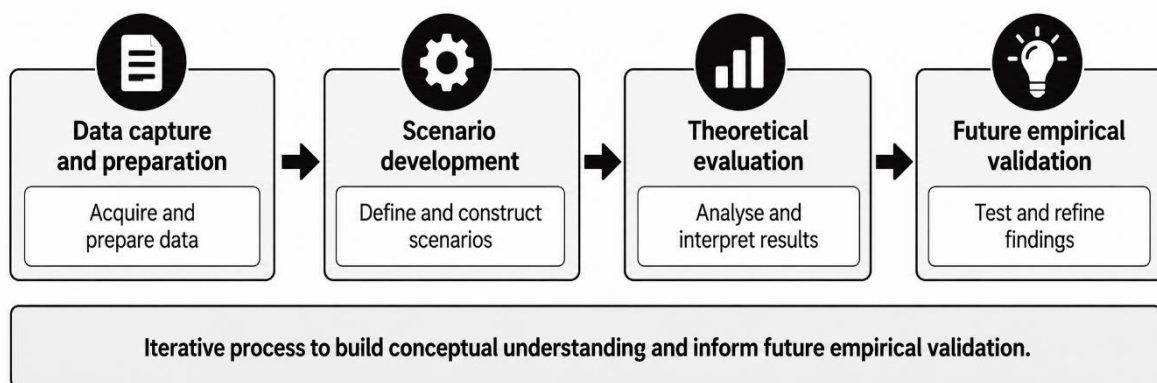


Figure 4. Conceptual validation logic.

Requirement	Architectural response	Validation status
DR1	L3 media-specific representation + replaceable L4 detector	Design coverage
DR2	L6 explanation report	Design coverage
DR3	L5 evidence correlation	Design + scenario coverage
DR4	L7 non-bypassable human review	Design + proposition
DR5	Separate future prediction/XAI metrics	Evaluation protocol
DR6	Replaceable L3/L4/L6 components	Design coverage
DR7	L7-L8 decision-support boundary	Governance coverage

DR8	L5 rights, territory, licence and chronology	Scenario coverage
DR9	Evidence can support or weaken a claim	Scenario + proposition

The validation does not establish detection accuracy, explanation fidelity, user comprehension or legal effectiveness. Future implementations should report precision, recall, F1, PR-AUC and class-specific errors for detection where appropriate. Explanation quality should be assessed separately through fidelity, stability, precision/coverage and clarity. Human-centred evaluation should examine comprehension, decision time, inter-reviewer agreement, calibration, appropriate reliance and the quality of overrides. Contemporary XAI research supports this multidimensional approach because perceived trust can increase without corresponding gains in understanding or decision quality [12], [4].

X. Discussion

A. The Case Record as the Unit of Accountability

The main theoretical move in XAI-IPFD is to shift attention from the prediction to the case record. A prediction answers what the model detected. Evidence asks what external facts support or weaken that suspicion. Explanation asks why the model behaved as it did. Human judgement determines the operational response. Auditability asks whether the reasoning can later be reconstructed. Treating these roles separately reduces the risk that a persuasive explanation is mistaken for proof or that a provenance record is mistaken for infringement.

This framing also clarifies the place of XAI. Explanation is necessary where complex models

influence consequential review, but it is not sufficient. A technically faithful explanation of an incorrect model remains incorrect, while a readable explanation cannot establish ownership. XAI-IPFD therefore embeds explainability inside an evidence-aware process rather than placing it at the end of a prediction pipeline.

B. Human Verification as Architecture, Not Decoration

Human oversight is often described as a governance safeguard added after a technical system has been designed. In XAI-IPFD, human verification has defined inputs, outputs and audit consequences. The reviewer receives the same prediction, uncertainty, evidence and explanation that are preserved in the final record. A challenge or dismissal is therefore not an exception to the system; it is part of the system's accountable operation.

This approach is aligned with recent human-centred XAI work that differentiates explanation needs across roles and levels of expertise. A developer may require detailed feature attributions, while a rights administrator may need concise evidence-linked reasons. A creator challenging a claim may need a different explanation again. The architecture therefore treats explanation design as stakeholder-dependent.

C. Comparative Positioning

Capability	Detection	Watermarking	Provenance	Conventional XAI	XAI-IPFD
Automated detection	Core	Limited	No	Core	Core
Provenance evidence	Limited	Strong when marked	Core	Limited	Integrated
Model explanation	Limited	No	No	Core	Core

Rights/territorial context	Rare	Limited	Possible	Rare	Explicit
Mandatory human verification	Rare	No	No	Variable	Explicit
Bidirectional protection	Rare	Limited	Limited	Rare	Explicit
Auditable case object	Variable	Partial	Strong record integrity	Variable	End-to-end

The table describes dominant functions in the reviewed streams rather than claiming that every system in a category has identical characteristics. XAI-IPFD's novelty lies in the disciplined

relationship among existing technical capabilities and the governance assumptions built into their interfaces.

D. Testable Research Propositions

Prop.	Layers	Testable proposition	Indicative measure
P1	L5-L6	Evidence-linked explanations will support better reviewer comprehension than model explanations presented without provenance and rights context.	Comprehension, evidence identification, decision quality
P2	L7	Mandatory human verification will reduce inappropriate conversion of ambiguous similarity into escalation.	Escalation error, override quality, calibration
P3	L6-L7	Explanation effectiveness will vary by reviewer role and expertise; no single explanation format will dominate across users.	Technique $\times$ stakeholder interaction
P4	L5	Rights-context modelling will improve handling of territorial, derivative-work and fraudulent-claim cases compared with content matching alone.	Contextual case accuracy
P5	L8	More complete audit records will improve contestability and reconstruction of the basis of a decision.	Reconstruction completeness, challenge quality

XI. Limitations and Research Agenda

Three limitations define the scope of the present claims. First, the evidence synthesis is integrative rather than an exhaustive multi-database systematic review. Second, XAI-IPFD is a conceptual architecture and does not yet report prototype performance. Third, institutional requirements will vary across rights holders, platforms, enforcement bodies and jurisdictions. Interface design, reviewer training, evidential rules and case-management

integration will therefore require contextual adaptation.

These limitations provide a staged validation programme. Stage one should expand the evidence base through Scopus, Web of Science, IEEE Xplore, ACM Digital Library, ScienceDirect and SpringerLink with transparent PRISMA reporting. Stage two should implement a bounded single-modality prototype, preferably audio or image, so that representation, detection, evidence and XAI

interfaces can be tested without unnecessary multimodal complexity. Stage three should evaluate human comprehension and reviewer decisions. Stage four should extend the architecture to multimodal and

institutional pilots. Stage five should translate validated controls into governance and deployment requirements.

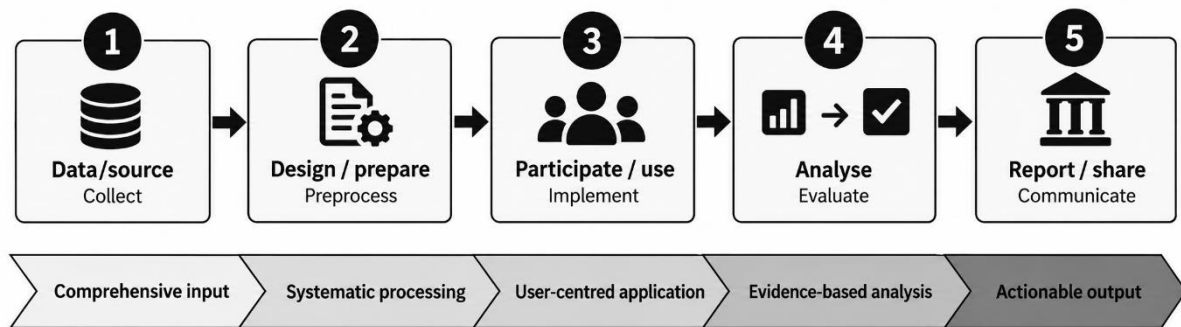


Figure 5. Staged research and validation programme.

A prototype study should compare an explainable configuration with the same detector operating without the explanation layer while holding training data and thresholds constant. Human studies should randomise explanation format where feasible and distinguish perceived trust from demonstrated understanding. The relevant question is not whether an explanation looks convincing, but whether it helps reviewers identify evidence, recognise uncertainty and appropriately challenge the model when it is wrong.

CONCLUSION

Digital piracy and intellectual property fraud occur within the same technological environment that has expanded creative participation and global distribution. Automated detection is necessary at platform scale, but detection alone is an incomplete institutional response. Creative-sector decisions also require evidence, rights context, explanation, human judgement and records that can be reconstructed and challenged.

XAI-IPFD contributes an eight-layer evidence-aware architecture that connects those functions while preserving their different epistemic roles. Its formal case object links prediction, uncertainty, evidence, rights context, explanation, human review and auditability. The framework is intentionally model-agnostic and can accommodate different media types, detectors and explanation techniques without

allowing any one component to become a substitute for evidence or human judgement.

The contribution remains conceptual rather than empirical. The article does not claim that LIME, SHAP or ANCHOR will achieve a particular performance level, that architectural coverage proves operational effectiveness, or that a technically supported case constitutes a legal determination. Instead, it provides a clearly specified object for subsequent testing. For the Orange Economy, the central proposition is that trustworthy IP protection requires more than finding similarity. It requires a transparent path from a machine-generated signal to evidence that authorised reviewers can inspect, explanations they can understand, decisions they can challenge and records institutions can audit.

REFERENCES

1. X. Chen, A. Yang, J. Weng, Y. Tong, C. Huang, and T. Li, "A blockchain-based copyright protection scheme with proactive defense," *IEEE Trans. Services Comput.*, vol. 16, no. 4, pp. 2316–2329, Jul.–Aug. 2023.
2. A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Q.*, vol. 28, no. 1, pp. 75–105, Mar. 2004, doi: 10.2307/25148625.
3. A. Hernandez-Suarez, G. Sanchez-Perez, L. K. Toscano-Medina, H. M. Perez-Meana, J. Portillo-Portillo, and J. Olivares-Mercado, "Methodological approach for identifying

- websites with infringing content via text transformers and dense neural networks," *Future Internet*, vol. 15, no. 12, p. 397, Dec. 2023, doi: 10.3390/fi15120397.
4. S. Hong and W. Park, "Developing user-centered system design guidelines for explainable AI: A systematic literature review," *Artif. Intell. Rev.*, vol. 58, art. no. 386, 2025, doi: 10.1007/s10462-025-11363-y.
5. R. Kumar, R. Tripathi, N. Marchang, G. Srivastava, and N. N. Xiong, "A secured distributed detection system based on IPFS and blockchain for industrial image and video data security," *J. Parallel Distrib. Comput.*, vol. 152, pp. 128–143, Jun. 2021, doi: 10.1016/j.jpdc.2021.02.022.
6. D. Li, H. Xin, and X. Jin, "Text feature-based copyright recognition method for comics," *Eng. Appl. Artif. Intell.*, vol. 132, art. no. 107925, Jun. 2024, doi: 10.1016/j.engappai.2024.107925.
7. C. F. F. Matias, "Access revisited: AI training at the intersection of copyright and cybercrime laws," *Comput. Law Secur. Rev.*, vol. 57, art. no. 106149, 2025, doi: 10.1016/j.clsr.2025.106149.
8. S. K. Padhi, A. Tiwari, and S. S. Ali, "Deep learning-based dual watermarking for image copyright protection and authentication," *IEEE Trans. Artif. Intell.*, vol. 5, no. 12, pp. 6134–6145, Dec. 2024, doi: 10.1109/TAI.2024.3485519.
9. J. Park, J. Kim, J. Seo, S. Kim, and J.-H. Lee, "DNN-based forensic watermark tracking system for realistic content copyright protection," *Electronics*, vol. 12, no. 3, p. 553, Jan. 2023, doi: 10.3390/electronics12030553.
10. J. P. Quintais, "Generative AI, copyright and the AI Act," *Comput. Law Secur. Rev.*, vol. 56, art. no. 106107, 2025, doi: 10.1016/j.clsr.2025.106107.
11. B. Raufi, C. Finnegan, and L. Longo, "A comparative analysis of SHAP, LIME, ANCHORS and DICE for interpreting a dense neural network in credit card fraud detection," in *Explainable Artificial Intelligence: xAI 2024*, vol. 2156, Springer, 2024, pp. 260–280, doi: 10.1007/978-3-031-63803-9_20.
12. M. Ridley, "Human-centered explainable artificial intelligence: An Annual Review of Information Science and Technology (ARIST) paper," *J. Assoc. Inf. Sci. Technol.*, vol. 76, no. 1, pp. 98–120, Jan. 2025, doi: 10.1002/asi.24889.
13. W. Saeed and C. Omlin, "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities," *Knowl.-Based Syst.*, vol. 263, art. no. 110273, Mar. 2023, doi: 10.1016/j.knosys.2023.110273.
14. A. Salih, Z. Raisi-Estabragh, I. Boscolo Galazzo, P. Radeva, S. E. Petersen, G. Menegaz, and K. Lekadir, "A perspective on explainable artificial intelligence methods: SHAP and LIME," *Adv. Intell. Syst.*, vol. 7, no. 1, art. no. 2400304, Jan. 2025, doi: 10.1002/aisy.20240030.
15. C. L. Saw and B. Z. Y. Tan, "Unpacking copyright infringement issues in the GenAI development lifecycle and a peek into the future," *Comput. Law Secur. Rev.*, vol. 58, art. no. 106163, 2025, doi: 10.1016/j.clsr.2025.106163.
16. S. Sharma, J. J. Zou, G. Fang, P. Shukla, and W. Cai, "A review of image watermarking for identity protection and verification," *Multimed. Tools Appl.*, vol. 83, no. 11, pp. 31829–31891, Apr. 2024, . <https://doi.org/10.1007/s11042-023-16843-3>
17. M. Szymanski, V. Vanden Abeele, and K. Verbert, "Disentangling stakeholder role and expertise in user-centered explainable AI," in *Proc. 33rd ACM Conf. User Model., Adapt. Personalization*, 2025, pp. 32–39, doi: 10.1145/3699682.3728351.

HOW TO CITE: Atiku Baba Shidawa^{*1}, Abednego Gambo Habu¹, Achi Aaron Unimke², Benjamin Opabunmi¹, Krulat Naandi Dariyem¹, Theresa Awulor³, XAI-IPFD: An Evidence-Aware Explainable Artificial Intelligence Architecture For Intellectual Property Fraud And Digital Piracy In The Orange Economy, *Int. J. Sci. R. Tech.*, 2026, 3 (10), 440-454. <https://doi.org/10.5281/zenodo.23210618>