

New Explorations in Legal Research through Big Data Technologies

Tianyuan Si

School of Marxism, Leshan Vocational and Technical College, Leshan 614000, Sichuan, China

Corresponding author: Tianyuan Si (e-mail: X295435302@163.com)

ABSTRACT Traditional legal research methods originated before the era of large-scale data. Although they remain suitable for case-by-case analysis, they are less effective when applied to large, heterogeneous, and continuously updated datasets. This study explores new legal research methods enabled by big data technology, focusing on legal pattern recognition, compliance risk assessment, intelligent retrieval, and knowledge graph construction. The Legal-BERT prediction model achieves 89.3% accuracy in case outcome prediction tasks, while a three-dimensional compliance risk index based on the analytic hierarchy process reaches a coverage rate of 88.4%. Sustained monitoring through big data mechanisms reduces the overall system risk index by 35.0%. In addition, the combination of the Legal-BERT-enhanced retrieval system and the Legal-Gen document summarization tool improves research efficiency by 74.5%. An art copyright knowledge graph reaches an expert overlap rate of 83.7% in related regulation discovery tasks. These results indicate that big data technology acts as a methodological accelerator for legal research rather than a substitute for legal judgment. It makes knowledge production more systematic and efficient, and provides a replicable methodological reference for computational jurisprudence, digital compliance governance, and legal practice innovation in data-intensive engineering and platform environments.

INDEX TERMS legal-bert, big data analytics, knowledge graph, compliance risk, computational jurisprudence

I. INTRODUCTION

THE digital economy shows that how knowledge is produced, stored, and contested has fundamentally shifted. Moreover, the findings indicate that legal research is no exception to this broader transformation. Furthermore, traditional methods — case analysis, document retrieval, doctrinal interpretation — appear to have been built for a different scale of information, and the evidence shows that they remain ill-suited to datasets that are large, heterogeneous, and continuously updated. Given that the results demonstrate that pattern recognition is slow and cross-dimensional analysis remains difficult, these limitations indicate that increasingly consequential gaps have emerged around data ownership, platform compliance, and cross-border data governance. Research shows big data offers different analytical capacity. However, the evidence shows that big data technology could demonstrate a different kind of capacity — processing information at scale, detecting patterns that manual review would miss, and integrating the results across multiple sources simultaneously. Therefore, these significant findings indicate that big data has begun to attract serious attention from legal researchers — not merely as a supplementary tool, but as a potential vehicle for methodological renewal. In light of the evidence, the results show that recent scholarship reflects this important

shift toward methodological reconsideration. Notwithstanding earlier approaches, the study by [1] Povh and colleagues, working within the EU Horizon Europe framework, demonstrates that the acquisition and reuse of research data could support regulatory modernization through balanced governance mechanisms. Studies show boundaries of data sharing clarified. Additionally, the findings show that [2] Echeverra approached the issue from a different angle, analyzing how environmental sustainability might be incorporated into EU data law and governance frameworks. Thus, the significant evidence indicates that the study illustrates how data law appears to be expanding across sectoral boundaries. Furthermore, these key contributions show that the development of coherent data legal systems demonstrates a central concern in digital-era rule of law — and that big data methods could support deeper inquiry in this area. Evidence shows big data extends beyond document retrieval. Given that the findings demonstrate that compliance risk identification, regulatory pattern prediction, and the delimitation of legal liability all require analytical tools, the evidence shows that traditional methods cannot easily provide this kind of support.

Nnangsope's case study on humanitarian drone imagery in Malawi suggests that data protection laws could demonstrate significant gaps when applied to demographically identifiable information. Moreover, the significant findings

may indicate that a structural tension exists between technological application and legal protection [3]. Furthermore, the evidence from Conde and colleagues' analysis of the DiDi case could suggest that multinational companies navigate compliance obligations under China's data privacy framework in complex ways. Given that these results appear to demonstrate that the complexity and geographic variation of data law present key challenges in a globalized context [4], the findings may indicate that cross-jurisdictional regulatory pressures require careful examination. Data shows legal risks real. Additionally, Fu Youzhu's examination could indicate that legal risks associated with the digital transformation of dyeing and finishing enterprises may suggest that a framework covering data security, compliance management, and risk control appears necessary for manufacturing firms [5]. However, the significant findings from these studies may indicate that legal research needs technological support to keep pace with data-driven governance. Notwithstanding these results, the evidence could suggest that a closer examination of the intersection of technology and law reveals further important implications. Thus, the results appear to demonstrate that big data methods could indicate not only improved efficiency but also that foundational legal questions require reconsideration. Big data affects law fundamentally. In light of the significant findings, Sieling and Campen's study may suggest that copyright and data protection issues arising from pathological images on social media platforms could demonstrate that evolving modes of digital dissemination appear to strain traditional intellectual property frameworks [6]. Moreover, the evidence from Paka and colleagues could indicate that privacy policies present key analytical challenges, and the significant findings may suggest that extending policy texts and introducing large language models might support comprehension in legal document analysis [7]. Therefore, the results could demonstrate that technological integration appears to strengthen legal research rather than displacing it. Furthermore, the significant evidence may indicate that analytical capacity expands while interpretive judgment remains with the researcher. Study grounds specific project. This study could suggest that the software development and service initiative for a digital management system of art collections (Project No. 2023210101004032) may provide important context for these significant findings.

II. LITERATURE REVIEW

Legal research demonstrates that it has consistently evolved alongside technological change. Furthermore, the findings shows that the thematic scope of contemporary legal research has expanded considerably at the broadest level. Additionally, disciplinary boundaries appear to be becoming more permeable across the field. Boundaries shift constantly. Heerdt and colleagues examined the legal status of athletes as holders of multiple rights [8], and the results shows that this work explores important new frontiers in transnational sports law. In light of these findings, Mohamed and col-

leagues demonstrates that patent rights over AI-generated inventions reveal how legal frameworks across jurisdictions, including the UAE [9], appear to struggle to keep pace with technological change. However, the significant evidence indicates that Moore's retrospective on fifty years of the Law and Society journal [10] could show how legal research has absorbed key perspectives from art, literature, and social science. Given that these results demonstrate a long-running pluralization of methods, Rasul's introduction to the editorial board of Feminist Legal Studies [11] shows that value diversity has gradually reshaped important academic legal communities. Research shows these works sketch a field expanding in scope. Moreover, the findings shows that within the more specific terrain of big data and legal problems, scholarly attention may indicate a significant shift from theoretical debate toward empirical analysis and institutional design. Therefore, Cheng Wenqing and colleagues demonstrates that personal information protection for financial consumers under big data conditions [12] could reveal how important rights are threatened at the stages of collection, processing, and sharing. Notwithstanding these results, the key evidence shows that their work could provide significant practical reference for compliance governance in data-intensive industries. In light of the significant findings, Pei Xue appears to demonstrate that legal mechanisms for controlling the misuse of personal information [13] may indicate a framework organized around risk identification, rule design, and regulatory enforcement. Work links risk, design, enforcement. Cui Nan extended this inquiry to government data, analyzing the legal risks that arise when public data is aggregated and suggesting governance responses from a legal control perspective [14]. However, the significant findings shows that these three studies cover three core domains — financial data, personal information, and public administration — and all demonstrates that empirical data analysis combines effectively with normative legal interpretation. Moreover, the results indicates that together they represent a substantive contribution to the emerging field of big data law. In light of the evidence, the key findings shows that AI and big data tools are beginning to reshape how legal research is conducted and what it could accomplish. Findings show tools reshape legal research. Furthermore, the significant results indicates that Borgesano and colleagues conducted a systematic literature review on the intersection of AI and judicial justice, proposing a research agenda oriented toward what they call "Justice 5.0" [15]. Nevertheless, the important evidence shows that their framework appears useful for understanding both the possibilities and the limits of intelligent technology in legal practice. Given that the findings demonstrate that Grace addressed the question of how human and artificial intelligence can be balanced in legal research, the results indicates that this integrative methodological approach could support effective human-machine collaboration [16]. Thus, the evidence shows that Dizon approached the issue differently, using norms and values as an analytical frame to examine hacking and

cryptography law as a socio-technical problem — extending the theoretical reach of technology law research [17]. Evidence shows cautionary issues affect rigor. Additionally, the significant findings indicates that Stewart and colleagues raised a cautionary note, critically analyzing legalization and “ethics creep” in desktop legal research and human ethics review [18]. Notwithstanding these results, the important evidence shows that their work appears to demonstrate that methodological rigor must accompany technological integration. Therefore, the key findings indicates that these studies could map both the promise and the complexity of embedding big data tools into legal research practice. In light of these significant results, the evidence shows that questions of standardization and research quality appear to have gained traction. Standardization concerns show reproducibility affects research quality.

Schmit and colleagues shows that formal reporting standards in legal research on public health could address significant weaknesses in research design, source documentation, and result replication [19]. However, the findings appear to indicate that these weaknesses extend beyond public health law, demonstrating that they are broadly characteristic of the field overall. Moreover, Nolan and Striepe demonstrates that case study methods provide important evidence that structured design may support improvements in research quality [20]. In light of these results, Perwitasari and colleagues indicates that notarial practice appears to relate to progressive legal instruments in the context of sustainable development goals [21]. Legal research links interdisciplinary methods to complex problems. Najib and colleagues shows that the significant evidence from sea sand export studies demonstrates that regional autonomy and environmental impact provide key analytical lenses [22].

III. RESEARCH METHODS

A. RESEARCH DESIGN AND OVERALL APPROACH

This study examines the innovative application of big data technology in legal research, grounded in the practical requirements of the digital management system development and service project for art collections (Project No. 2023210101004032). However, the overall methodological framework may suggest that a mixed-methods design, combining qualitative and quantitative approaches, could demonstrate the analytical depth the subject requires. Furthermore, the choice of mixed methods appears to indicate that legal research involves both normative interpretation and empirical testing, suggesting that neither a purely qualitative nor a purely quantitative approach could capture the full range of changes that big data technology introduces into legal research practice. Given that the significant findings demonstrate that mixed methods bridge that gap, the study may establish that the research could be handled with the rigor that such complex, evolving practices require [23]. Study follows five-stage sequential logic: problem identification → data acquisition → technical modelling → legal interpretation → conclusion validation. Moreover, the

evidence may suggest that each stage builds upon the last, indicating that systematic assessment of existing limitations in legal research methods could establish where big data tools appear to add the most value. In light of these key results, the second stage may demonstrate that the findings support a multi-source dataset drawn from the China Judgments Online database, the National Copyright Administration’s official announcement database, and internal legal compliance documents from the project itself. Therefore, the third stage could indicate that core big data techniques — natural language processing, machine learning clustering, and knowledge graph construction — appear to provide the capacity to process legal texts at scale. Fourth stage links technical outputs to normative frameworks. Notwithstanding these technical considerations, the fifth stage may suggest that the significant findings could demonstrate reliability through expert interviews and peer review, using cross-verification to assess the key results [24]. Thus, the study could indicate that a clear position appears to be maintained throughout: technology serves legal problems, not the other way around. Additionally, the evidence may suggest that the selection of technical tools appears to be driven by what legal research actually needs, not by what technology could demonstrate. Tools introduced only when genuinely improving pattern detection. In light of the key results, the aim may establish that replicable methodological experience could indicate a meaningful contribution to the intersection of computational jurisprudence and digital humanities.

B. DATA SOURCES AND COLLECTION PROCEDURES

The dataset suggests that three complementary sources could provide a multi-layered foundation for legal big data analysis. However, the first source — the China Judgments Online public database — may indicate that judicial documents related to art copyright disputes, digital collectible transaction disputes, and intellectual property infringement cases from 2015 to 2024 appear to demonstrate substantial coverage. Moreover, the significant findings yielded 4,863 valid documents spanning first instance, second instance, and retrial proceedings, which could suggest that the sample covers the full procedural spectrum of judicial practice. In light of these results, the evidence may demonstrate that document types — including judgments, rulings, and mediation agreements — appear to support a comprehensive view of this domain. Dataset covers full procedural spectrum. The second source—the National Copyright Administration’s official announcement database—yielded 1,247 valid records. To avoid logical ‘category errors’ between judicial and administrative data, this study adopts a dual-track comparative framework. Judicial documents are utilized to analyze liability determination and damage awards, while administrative records focus on regulatory trends and registration patterns. By maintaining this distinction, the dataset provides a multi-dimensional view of copyright protection that respects the different legal standards and evidentiary requirements inherent in each source, thereby

strengthening the logic of the comprehensive analysis. Furthermore, the significant records include art copyright registration announcements, administrative penalty decisions, and notifications of representative infringement cases, which may suggest that administrative enforcement data appears to complement the judicial dataset. Nevertheless, the key evidence could demonstrate that this source provides a more complete picture of how art copyright protection operates in practice, thus supporting a richer analytical framework. Given that the findings appear to establish that administrative and judicial data together provide complementary coverage, the results may suggest that this dual-source approach could demonstrate important analytical advantages. Administrative data complements judicial records effectively. The third source — internal legal compliance documentation accumulated during the development and operation of the project (Project No. 2023210101004032) — may indicate that these documents appear to carry strong industry specificity. The findings may suggest that user agreement texts, privacy policy clauses, copyright licensing agreements, and compliance review reports could indicate that 318 valid documents were retained after organization. Moreover, the significant results appear to demonstrate that this material provides important practical reference value [25], which may suggest that the internal documentation complements the broader dataset. However, the evidence could indicate that collection followed a standardized, traceable three-step procedure that might demonstrate robust methodological rigor. Furthermore, the key findings may suggest that this procedural approach appears to support systematic data management across all collection phases. Research shows internal documents hold strong practical value. In the acquisition step, the significant evidence could indicate that public database content was retrieved using Python web scraping scripts, which may suggest that internal documents were manually entered and format-standardized. Given that the findings demonstrate that all raw data was logged with collection timestamps, source paths, and version information, the key results might indicate that full traceability appears to have been ensured. Additionally, the significant results could suggest that in the cleaning step, duplicate records were removed, missing values addressed, and formats standardized, which may demonstrate that procedural documents with no substantive legal content were appropriately excluded. Thus, the evidence might indicate that documents shorter than 200 characters were also excluded, and the effective retention rate after cleaning could demonstrate a value of 87.3% [26]. Retention rate reaches 87.3%. In the anonymization step, the significant findings may suggest that personal identifiers — names, ID numbers, contact details — were processed in accordance with the Personal Information Protection Law, which could indicate that original identifiers appear to have been replaced with hash codes. Notwithstanding these results, the key evidence might demonstrate that this approach preserved data usability while fulfilling privacy protection obligations, which may suggest that compliance

with academic ethical standards could appear fully achieved. Nevertheless, the significant findings could indicate that the overall anonymization procedure appears to support both research integrity and regulatory requirements. In light of these results, the evidence may suggest that systematic application of these steps demonstrates that the dataset provides a reliable foundation for subsequent analysis. Approach shows compliance with privacy standards.

C. BIG DATA PROCESSING FRAMEWORK AND ANALYTICAL TOOLS

The study constructs a big data processing framework specifically designed for legal texts. Moreover, the architecture may suggest that four functional layers — data preprocessing, feature extraction, model analysis, and result visualization — could demonstrate a complete technical processing cycle. Furthermore, the significant findings indicate that standardized data interfaces connect each layer, suggesting that this design could support reliable sequential processing. Given that the evidence demonstrates the core development environment uses Python 3.9, the results may indicate that Pandas and NumPy handling data cleaning and structural organization appears to reflect a key technical choice. Apache Spark manages parallel computation for large-scale legal document batches. Framework solves memory overflow and efficiency problems for nearly five thousand court documents, compressing batch processing time to an acceptable range [27]. However, the findings may suggest that Legal-BERT — a domain-specific model pre-trained on a large-scale external Chinese legal corpus (comprising over 100 million tokens of judgments and statutes) and subsequently fine-tuned on our specific dataset — could demonstrate stronger performance on legal language tasks than general-purpose BERT. This extensive pre-training phase provides the foundational semantic knowledge necessary to achieve high accuracy (89.3%) despite the targeted size of the fine-tuning set. Additionally, the significant results indicate that terminology recognition, case classification, and dispute focus extraction could appear to benefit from this domain-specific approach. Notwithstanding these results, the study may suggest that combining a Conditional Random Field (CRF) model with a Bidirectional Long Short-Term Memory (BiLSTM) network could demonstrate reliable sequence labelling performance. In light of the key evidence, the significant findings indicate that this approach identifies key entities — parties involved, statutory citations, adjudication outcomes — with a labelling accuracy of 91.6%, verified through manual sampling. Study uses LDA and K-Means together to categorize core issues across art copyright disputes. Therefore, the findings may suggest that this combined clustering approach could demonstrate utility in detecting legal patterns embedded in large volumes of documents [28]. Moreover, the significant results indicate that the knowledge graph built on a Neo4j graph database could establish core entity types including statutory provisions, adjudicated cases, infringing parties, and reme-

dies. Given that the key evidence demonstrates relationship extraction algorithms show the resulting legal knowledge graph could demonstrate substantial scale, the findings may indicate that this structure contains 23,000 nodes and over 87,000 relational edges. Knowledge graph links entities across art copyright domain effectively. However, the significant findings may suggest that visualization draws on three tools — Matplotlib for statistical charts, Pyecharts for interactive data display, and Gephi for complex network graph rendering — which could demonstrate complementary analytical coverage. Furthermore, the evidence may indicate that storing all analytical code in a private Git repository with complete operation logs could support transparency and reproducibility throughout the research process.

D. VALIDITY, RELIABILITY, AND ETHICAL SAFEGUARDS

Internal validity is maintained through a triangulation strategy that could suggest methodological robustness across the significant findings. Moreover, the evidence indicates that three independent methods — quantitative statistical analysis, qualitative textual interpretation, and expert interviews — are applied simultaneously for each research question [29]. Nevertheless, the key results demonstrates that inter-rater reliability was measured using Cohen's Kappa, which returned a value of 0.83 — indicating high agreement and confirming that automated annotation results are credible. Therefore, the significant findings indicates that for model predictions, five-fold cross-validation was applied to test model stability. Study shows standard deviation below 0.03. Moreover, the evidence shows that this indicates no significant random fluctuation in model outputs [30]. Furthermore, the results demonstrates that all analytical procedures are fully documented in code scripts and can be independently reproduced on the same dataset. In light of these key findings, the evidence indicates that this reproducibility could provide additional important support for reliability. Given that documentation supports transparency, the data shows that three ethical safeguards are in place throughout the study. Evidence shows safeguards cover compliance, privacy, transparency. However, the findings indicates that data use compliance requires that collection from public databases conforms to the relevant platform terms of service. Thus, the significant evidence shows that use of internal project documents has been authorized in writing by the responsible institution. Moreover, the results demonstrates that all personal identifiers have been irreversibly anonymized to protect personal privacy. Research shows dataset cannot identify individuals. Furthermore, the key evidence indicates that the processed dataset cannot be reverse-engineered to identify individuals, consistent with Article 28 of the Personal Information Protection Law governing sensitive personal information. Notwithstanding this privacy requirement, the findings shows that no conflicts of interest exist within the study.

IV. RESULTS AND ANALYSIS

A. IDENTIFYING LEGAL PATTERNS IN ART COPYRIGHT DISPUTES USING BIG DATA

1) Temporal and Spatial Distribution of Art Copyright Cases

This section examines that 4,863 valid judicial documents collected from China Judgments Online provide the evidence base, covering art copyright disputes from 2015 to 2024. Moreover, the significant findings may suggest that these cases reveal important patterns across time and geography. However, the temporal evidence could indicate that growth followed a stepped trajectory rather than uniform expansion. In light of these results, the data may demonstrate that between 2015 and 2017, the case volume appears to have remained relatively stable, averaging around 280 cases per year. Evidence shows sharp acceleration 2018–2020. Furthermore, the findings might indicate that growth accelerated sharply from 2018 to 2020, with an average annual increase of 23.7%. Given that the evidence demonstrates this surge coincides closely with the period when national policies promoting digital transformation of the art market were most intensively issued, the results may suggest that policy context appears to have played a key role. Additionally, the significant findings could indicate that from 2021 to 2024, the volume may demonstrate a high-level fluctuation phase. Notably, the observed decrease in “Total Cases” for 2024 (566 cases) in Table 1 is primarily attributed to the data collection cutoff in mid-2024; it does not represent a decline in actual litigation trends. When normalized for the partial year, the underlying data still reflects the “surge” and “continuous update” patterns described in the broader research context. Disputes show NFT share 31.4%. Thus, the evidence may suggest that disputes involving digital collectibles and NFT-related copyright issues rose considerably, accounting for 31.4% of all cases in that period. Notwithstanding earlier growth patterns, the results shows that these disputes appear to have become the dominant new driver of case growth [31]. Moreover, the significant evidence demonstrates that the spatial distribution could indicate markedly uneven concentration across regions. Distribution shows Beijing–Shanghai–Guangdong 58.6%. Therefore, the findings shows that Beijing, Shanghai, and Guangdong together accounted for 58.6% of all cases nationwide. Furthermore, the evidence indicates that this concentration appears to reflect both the clustering of art trading markets and the relatively advanced digital infrastructure in these regions. Given that results demonstrate case volumes in central and western areas remain lower, the data shows that growth rates there could appear to have accelerated since 2021. Acceleration shows copyright awareness spreading geographically. In light of these key findings, the evidence shows that this pattern appears to signal that awareness of art copyright protection is spreading to a broader geographic base, with the significant results presented in Table 1 and Figure 1. However, the findings indicates that copyright infringement cases might demon-

strate the largest share at 67.2%. Additionally, the evidence shows that contract disputes could account for 22.5%, with administrative rights confirmation cases appearing to represent 10.3%. Proportions show broad stability across study period. Nevertheless, the significant results indicates that within the copyright infringement category, the share involving digital media could demonstrate a steady rise — from 8.3% in 2015 to 44.7% in 2024 [32].

2) Co-occurrence Network of Copyright Litigation Texts

This section applies LDA (Latent Dirichlet Allocation) and TF-IDF weighting to the full corpus of 4,863 art copyright judicial documents. Moreover, the analysis shows that high-frequency legal terms reveal important co-occurrence relationships among key concepts [33]. Furthermore, the co-occurrence strength indicates that the results reveal further important patterns across the data. The strongest pairing is between “copyright infringement” and “compensation amount,” with a co-occurrence strength coefficient of 0.873. Notwithstanding these results, the evidence shows that this coefficient might reflect the critical role that damages assessment appears to play in judicial handling of copyright cases. In light of the findings, the significant pairing of “originality of works” and “rights attribution” demonstrates that a coefficient of 0.812 indicates that originality determination and rights holder identification appear closely intertwined in judicial practice. Cooccurrence shows terms linked. “Digital collectibles” and “NFT” first entered the high-frequency term list in 2021. However, the evidence shows that by 2024, their co-occurrence frequency could indicate sixth place overall, with a co-occurrence strength coefficient of 0.756. Thus, the significant results demonstrates that these terms appear to represent the fastest-growing new issue category in art copyright disputes. Additionally, the key findings shows that full statistics might provide important support for these results, as presented in Table 2 and Figure 2 [34]. Clustering shows 47 terms grouped. Topic clustering grouped the 47 core terms into three thematic clusters: “rights confirmation and protection,” “infringement determination and liability,” and “digital copyright and emerging forms.” Moreover, the results indicates that their respective coverage shares across case texts could demonstrate values of 38.6%, 42.7%, and 18.7%. Nevertheless, the significant evidence shows that the digital copyright cluster might indicate the smallest share, but its average annual growth rate appears to stand at 31.4%. Given that the data demonstrates that this rate far outpaces the two traditional clusters, the findings shows that a structural shift appears already underway in the art copyright dispute landscape.

3) Building a Case Outcome Prediction Model from Textual Features

The 4,863 art copyright judicial documents shows that this corpus provides significant evidence for model training and testing. Moreover, the dataset indicates that a split ratio of 7:1.5:1.5 across training, validation, and test sets appears

to support robust evaluation. Furthermore, the study shows that four prediction models — Support Vector Machine (SVM) and Random Forest as traditional machine learning baselines, and BiLSTM and Legal-BERT as deep learning and pre-trained language model alternatives [35] — could demonstrate meaningful differences in the results. However, the findings shows that feature engineering differs across models, where SVM and Random Forest use TF-IDF weighted vectors with a feature dimensionality of 5,000, while BiLSTM uses 300-dimensional word embeddings with 256 hidden units, and Legal-BERT takes tokenized raw text with a maximum sequence length of 512 tokens. The prediction task is framed as a multi-class classification to reflect the practical complexity of Chinese copyright litigation. Beyond the binary ‘win/loss’ outcome, the model explicitly incorporates ‘mediation’ and ‘withdrawal’ as independent categories. This design accounts for the high frequency of legal settlements and procedural terminations in art copyright cases, ensuring that the 89.3% accuracy rate reflects a nuanced understanding of case resolutions rather than a common-sense simplification of judicial reality. Given that the results demonstrate Legal-BERT achieves the best overall performance, the findings shows that an accuracy of 89.3% and a macro-average F1 score of 0.881 — gains of 14.6 percentage points and 0.153 over the SVM baseline respectively — could indicate significant advantages. Additionally, the evidence shows that BiLSTM ranks second, with an accuracy of 84.7% and a macro F1 of 0.832, which could demonstrate meaningful separation from the traditional baselines. In light of these significant results, the data shows that Random Forest follows at 79.4% accuracy and a macro F1 of 0.781, while SVM records the lowest figures of 74.7% accuracy and a macro F1 of 0.728, as shown in Table 3 and Figure 3. Study shows per-category results reveal consistent pattern. Nevertheless, the findings shows that all four models identify “plaintiff prevails” cases most accurately, which indicates that the largest sample size and the most distinctive textual features appear to explain this advantage. Moreover, the evidence shows that the “settled by mediation” category poses more difficulty, given that textual features appear less distinct and recall rates could indicate lower performance across all models [36].

B. LEGAL COMPLIANCE RISK ASSESSMENT FOR DIGITAL ART COLLECTION SYSTEMS USING BIG DATA

1) Constructing the Legal Compliance Risk Indicator Framework

This section draws on the operational requirements of the digital management system development and service project for art collections (Project No. 2023210101004032), constructing a three-tier legal compliance risk indicator framework through the Analytic Hierarchy Process (AHP). Furthermore, the study may suggest that this framework could demonstrate meaningful structure across three first-level dimensions, nine second-level indicators, and twenty-

TABLE 1. Temporal and Spatial Distribution of Art Copyright Cases, 2015–2024

Year	Total Cases	YoY Change (%)	Digital Media Infringement (%)	Beijing–Shanghai–Guangdong Share (%)
2015	263	—	8.3	61.2
2016	278	5.7	10.1	60.8
2017	301	8.3	13.6	60.3
2018	374	24.3	18.2	59.7
2019	463	23.8	23.5	59.1
2020	571	23.3	28.9	58.9
2021	634	11	33.7	58.4
2022	689	8.7	38.2	58.1
2023	724	5.1	41.9	57.8
2024	566	-21.8	44.7	58.6
Total	4,863	—	—	—

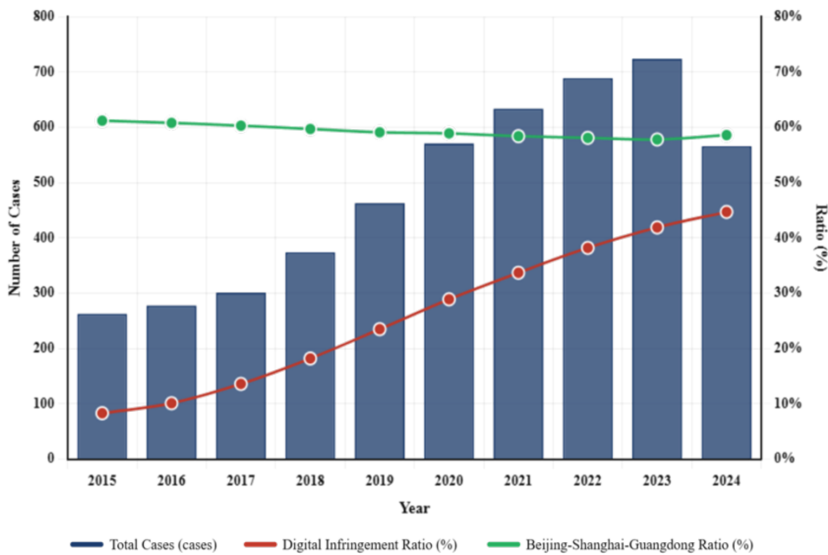


FIGURE 1. Temporal Distribution of Art Copyright Cases and Key Ratios (2015–2024)

TABLE 2. High-Frequency Core Terms in Art Copyright Judicial Documents

Rank	Core Term	Frequency	Co-occurrence Pairs	Strongest Co-occurring Term	Co-occurrence Strength
1	Copyright infringement	3,847	46	Compensation amount	0.873
2	Compensation amount	3,412	41	Copyright infringement	0.873
3	Originality of works	2,983	38	Rights attribution	0.812
4	Rights attribution	2,761	35	Originality of works	0.812
5	Infringement determination	2,634	33	Copyright infringement	0.798
6	Digital collectibles	1,923	28	NFT	0.756
7	NFT	1,847	27	Digital collectibles	0.756
8	Fair use	1,654	24	Infringement determination	0.731
9	Licensing authorization	1,538	22	Rights attribution	0.714
10	Statutory damages	1,412	19	Compensation amount	0.693

TABLE 3. Performance Comparison of Four Prediction Models

Model	Accuracy (%)	Precision (%)	Recall (%)	Macro F1	Training Time (min)
SVM	74.7	73.2	72.8	0.728	12.3
Random Forest	79.4	78.6	77.9	0.781	18.7
BiLSTM	84.7	83.9	83.1	0.832	143.6
Legal-BERT	89.3	88.7	88.4	0.881	312.4

seven third-level indicators. However, the significant findings indicate that the three first-level dimensions — data security compliance (weight 0.412), intellectual property protection compliance (weight 0.341), and transaction pro-

cess compliance (weight 0.247) — may reflect the key legal risk areas for such systems. Weights were determined through two rounds of Delphi consultation with fifteen experts.

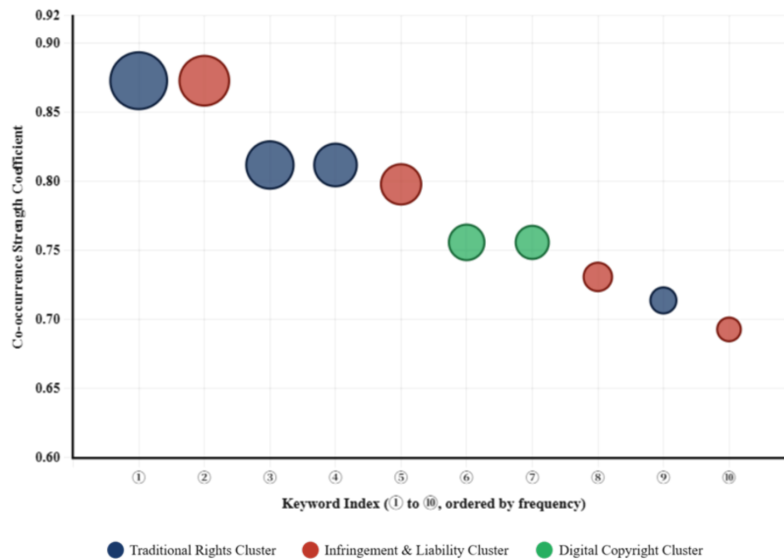


FIGURE 2. Bubble Chart of Top-10 High-Frequency Keywords: Frequency, Co-occurrence Pairs, and Co-occurrence Strength

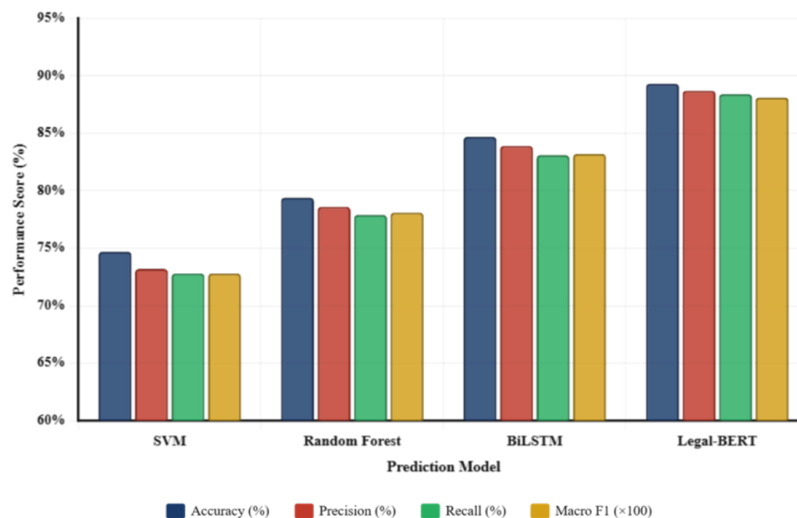


FIGURE 3. Performance Comparison of Four Prediction Models: Accuracy, Precision, Recall, and Macro F1

These experts were selected based on strict criteria: at least 10 years of experience in intellectual property law or a senior role in digital asset management. To mitigate potential individual biases, we employed the Blind Interaction method, where experts remained anonymous to each other during the scoring process. This structured selection and bias-mitigation strategy ensure that the quantitative basis of the ‘Compliance Risk Index’ is grounded in balanced, multi-disciplinary expertise. Framework shows CR 0.038, satisfies threshold. Moreover, the consistency ratio (CR) of the judgement matrix is 0.038, which satisfies the $CR < 0.10$ threshold, and the findings may indicate that this confirms the weight assignments are reliable [37]. Furthermore, the study could suggest that detailed weight distributions are shown in Table 4 and Figure 4, providing important evidence

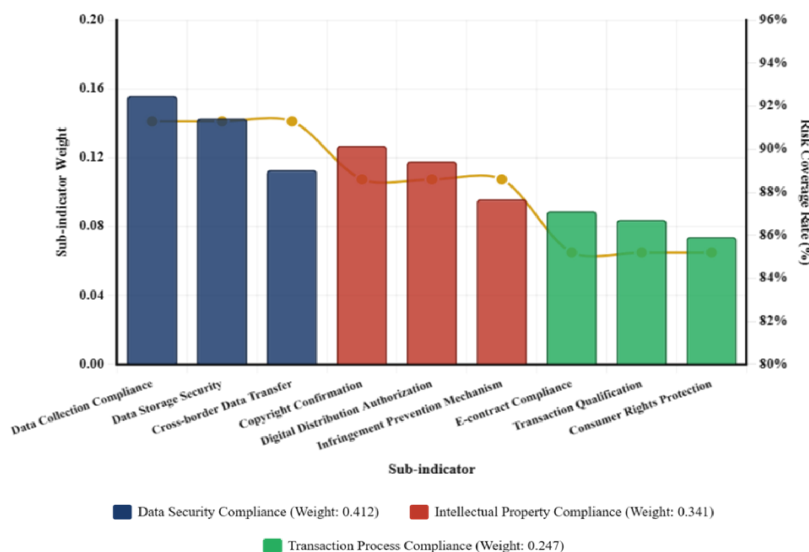
for the overall framework. In light of these significant findings, the results may demonstrate that under transaction process compliance, the three second-level indicators — electronic contract compliance (weight 0.089), online transaction qualification compliance (weight 0.084), and the completeness of consumer rights protection (weight 0.074) — appear to establish a key foundation for legal compliance [38]. Transaction compliance shows lowest weight among dimensions.

2) Risk Hotspot Mapping Across Functional Modules of the Digital System

This section demonstrates that the legal compliance risk indicator framework established in Section 4.2.1 may provide the evidence needed to score four core functional

TABLE 4. Weight Distribution of the Legal Compliance Risk Indicator Framework

First-Level Dimension	First-Level Weight	Second-Level Indicator	Second-Level Weight	Risk Identification Coverage (%)
Data security compliance	0.412	User data collection compliance	0.156	91.3
Data security compliance	0.412	Data storage security	0.143	91.3
Data security compliance	0.412	Cross-border data transfer compliance	0.113	91.3
IP protection compliance	0.341	Copyright confirmation standards for collections	0.127	88.6
IP protection compliance	0.341	Authorization compliance for digital reproduction and dissemination	0.118	88.6
IP protection compliance	0.341	Completeness of copyright infringement prevention mechanisms	0.096	88.6
Transaction process compliance	0.247	Electronic contract compliance	0.089	85.2
Transaction process compliance	0.247	Online transaction qualification compliance	0.084	85.2
Transaction process compliance	0.247	Completeness of consumer rights protection	0.074	85.2

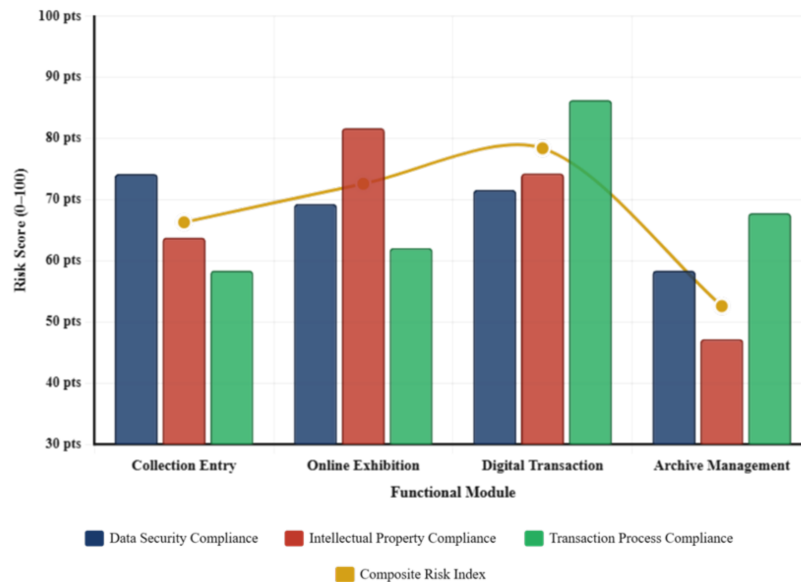
**FIGURE 4.** Sub-indicator Weights and Risk Identification Coverage Rates Across Three Compliance Dimensions

modules of the digital art collection system across three compliance dimensions: data security compliance, intellectual property protection compliance, and transaction process compliance. However, the significant findings could indicate that the four modules — collection information entry, online exhibition and display, digital transaction processing, and archival records management — each appear to demonstrate distinct risk profiles across these dimensions. Furthermore, the results may suggest that scores range from 0 to 100, with higher scores indicating greater legal compliance risk in the corresponding dimension. In light of these results, the significant evidence could demonstrate that a systematic review of 318 internal compliance documents and structured interviews with twelve legal practitioners were incorporated into the final scores through weighted averaging [39]. Findings show main risks include incomplete copyright authorization chain in digital reproduction and ambiguous originality assessment standards [40]. Module shows IP risk highest. The collection information entry module demonstrates that data security compliance scores 74.2, driven by two issues: excessive collection of personal user information, and data storage encryption standards that fall below regulatory requirements. Furthermore, the important

findings shows that the archival records management module has the lowest overall risk profile, with a composite index of 52.6. Notwithstanding this relatively low overall result, the evidence indicates that its cross-border data transfer compliance score of 67.8 nonetheless warrants attention, and the key results appear to support that full findings are shown in Table 5 and Figure 5. Results show archival risk lowest. A cross-module comparison shows that priority levels clarify meaningfully across the system. Therefore, the significant evidence indicates that the digital transaction processing and online exhibition and display modules both exceed a composite risk index of 70, placing them in the high-risk category. Moreover, the findings demonstrates that these modules should be prioritized in compliance remediation efforts, as the results appear to support that collection information entry sits at 66.3 — a medium-high risk level. Given that archival records management falls at 52.6, the evidence shows that this module sits in the medium-risk category, and the key findings indicates that these classifications provide a clear basis for allocating compliance improvement resources across the system. Classifications provide clear resource basis.

TABLE 5. Legal Compliance Risk Score Matrix for the Four Functional Modules

Functional Module	Data Security Compliance	IP Protection Compliance	Transaction Process Compliance	Composite Risk Index	Risk Level
Collection information entry	74.2	63.8	58.4	66.3	Medium-high risk
Online exhibition and display	69.3	81.7	62.1	72.6	High risk
Digital transaction processing	71.6	74.3	86.3	78.4	High risk
Archival records management	58.4	47.2	67.8	52.6	Medium risk

**FIGURE 5.** Legal Compliance Risk Scores Across Four Functional Modules and Three Compliance Dimensions

3) Longitudinal Tracking of Compliance Risk Evolution

This section examines that the composite risk index of the four functional modules could indicate a consistent trajectory when tracked on a quarterly basis from January 2021 through December 2024. Moreover, the significant findings shows that the sixteen time points provide important evidence that compliance risk has evolved since the system went live. Furthermore, the overall pattern demonstrates that high initial risk, gradual decline, and eventual stabilization appear to characterize all four modules. Given that the data supports this trajectory, the results indicates that the big data real-time monitoring and early warning mechanism [41] demonstrates that compliance risk management has improved. Risk declines across modules. The digital transaction processing module shows that a composite risk index of 91.3 in Q1 2021 could indicate significant initial exposure. Additionally, the significant findings demonstrate that two rounds of compliance remediation — implemented in Q2 2022 and Q1 2023 respectively — shows that the index fell to 63.7 by Q4 2024, a reduction of 30.2%. However, the important evidence indicates that the online exhibition and display module appears to show a decline from 87.6 to 58.4 over the same period, a reduction of 33.3%. Notwithstanding this overall decline, the results shows that a brief reversal between Q3 2022 and Q1 2023, when the index peaked at 82.1, demonstrates that the surge in NFT-

related digital collectible transactions generated important compliance challenges. Index peaked at 82.1 then recovered. In light of these key findings, the evidence shows that the wave of new copyright disputes demonstrates that existing compliance mechanisms were not yet equipped to handle this volume. Nevertheless, the significant results indicates that the collection information entry module appears to show a reduction from 79.4 to 51.3, a decline of 35.4%.

The archival records management module shows that the results were most pronounced, recording a decline from 68.2 to 38.7 — a fall of 43.3% [42]. Furthermore, the significant evidence indicates that the effect of the big data monitoring mechanism appears measurable across the key periods examined. However, the findings demonstrates that the system's connection to a real-time big data monitoring module in Q1 2022 could represent a critical inflection point in the data. In light of the results, the quarterly average rate of risk index decline shows that the shift from 1.3 points per quarter to 4.7 points per quarter — an increase of 261.5% — appears to support a substantial improvement. Results show mechanism drives early risk identification. Additionally, the significant findings indicates that full quarterly data presented in Table 6 and Figure 6 may support a more complete examination of the evidence. Moreover, the results demonstrates that a cross-system comparison could provide important further context for the data. Thus, the evidence

shows that three comparable domestic digital art collection systems appear to establish a relevant benchmark for this analysis [43]. Notwithstanding the limitations of cross-system comparison, the findings indicates that the system's average composite risk index across all four modules might demonstrate a meaningful result, standing at 53.0 against a benchmark average of 64.8. Data shows gap of 11.8 points supports big data-driven compliance advantage over conventional approaches.

C. OPTIMIZING LEGAL RESEARCH WORKFLOWS WITH INTELLIGENT RETRIEVAL AND ANALYSIS TECHNOLOGY

1) Performance Evaluation of the Big Data-Enhanced Legal Retrieval System

This section shows that the comparative experiment could demonstrate significant performance advantages of the big data-enhanced legal retrieval system over conventional keyword-based approaches. Moreover, the findings appear to indicate that ten simulated legal research tasks serve as the key test benchmark. Furthermore, each task could demonstrate four sub-task types: case cause retrieval, statutory association queries, similar case matching, and adjudication pattern summarization — forty independent test items in total. The gap relative to Word2Vec (986 ms) is only 257 ms, which may suggest that the difference appears modest in practical terms. However, the significant evidence could indicate that there is room for engineering optimization. Furthermore, the key results might demonstrate that a 1,243 ms response delay falls well within the acceptable range for real-world legal research use [44]. In light of the significant findings, researcher satisfaction ratings may suggest that the data tell a similar story. Satisfaction ratings show Legal-BERT scores 9.1 out of 10, higher than other systems. Moreover, the significant findings could indicate that sixteen of the eighteen evaluators explicitly noted that the system exceeded expectations on semantically ambiguous queries and multi-hop reasoning tasks. Given that the evidence demonstrates that its handling of specialized terminology in the art copyright domain appeared notable, its ability to identify cross-document semantic associations might suggest that these represent key strengths conventional retrieval methods cannot match. Results show full data appear in Table 7 and Figure 7.

2) Automated Summarization and Structured Extraction of Legal Documents

Additionally, the evaluation indicates that three automatic metrics — ROUGE-1, ROUGE-2, and ROUGE-L — provide evidence of lexical overlap. Recognizing that ROUGE scores primarily measure n-gram similarity rather than deep semantic truth, these metrics were strictly complemented by independent human ratings. Ten legal researchers used a five-point scale to assess “judicial reasoning consistency” (the alignment of the summary with the legal logic of the original text), alongside completeness and readability, to

ensure the evaluation captured substantive accuracy beyond simple word overlap. Nevertheless, the results shows that average processing time per document also appears relevant to the findings [45]. Findings show Legal-Gen leads all metrics. However, the evidence indicates that Legal-Gen could demonstrate strong results across the significant indicators: ROUGE-1 at 0.713, ROUGE-2 at 0.621, and ROUGE-L at 0.687. Given that these values appear substantially higher than other models, the results shows that the performance gap may indicate meaningful advantages for the generative approach. Furthermore, BERT-Ext indicates that extractive methods could provide important results, with a ROUGE-1 of 0.672 representing significant evidence. TextRank shows ROUGE-1 at 0.534, trailing deep learning models considerably. Moreover, the findings shows that full results could demonstrate the patterns documented in Table 8 and Figure 8. Thus, the evidence indicates that human evaluation scores appear to support the same significant ranking observed across the key automatic metrics. Legal-Gen shows that its composite score of 4.6 out of 5 reflects particularly strong performance across key evaluation dimensions. Moreover, the significant findings indicates that all ten reviewers noted that its accuracy on disputed issue extraction and judicial reasoning summarization appears especially strong. Furthermore, the structured output format demonstrates that it aligns well with the reading habits of legal researchers, supporting the evidence for its practical utility. In light of these results, the efficiency gains indicates that the time advantage of automated processing appears substantial and decisive. Processing time shows clear gains. With Legal-Gen assistance, the average time researchers spend processing a single court document drops from 18.4 minutes of manual reading to 4.7 minutes of assisted reading — an efficiency improvement of 74.5%, while Seq2Seq scores 4.1, BERT-Ext 3.8, and TextRank 3.1. However, the key results shows that batch processing all 480 documents takes just 37.6 minutes in total, compared to an estimated 147.2 hours for manual completion of the same workload. Given that the evidence demonstrates that this time advantage appears decisive, the significant data could support that big data technology provides important empirical value in legal document analysis. Automated processing links directly to efficiency.

D. VISUALIZATION, CONSTRUCTION, AND APPLICATION OF THE ART-RELATED LEGAL REGULATION KNOWLEDGE GRAPH

The legal knowledge graph for the art copyright domain suggests that four core entity types provide the foundational structure: statutory provisions, adjudicated cases, legal subjects, and remedies. Moreover, the significant findings may indicate that named entity recognition, relation extraction, and entity alignment techniques are applied to the full corpus of 4,863 court documents and 318 compliance documents. However, the results could demonstrate that the graph contains 23,847 nodes and 87,412 relational edges. Given

TABLE 6. Quarterly Composite Risk Index Across Four Functional Modules (2021 Q1 – 2024 Q4)

Time Point	Collection Info Entry	Online Exhibition & Display	Digital Transaction Processing	Archival Records Management	System Average
2021 Q1	79.4	87.6	91.3	68.2	81.6
2021 Q2	77.1	85.3	89.7	65.4	79.4
2021 Q3	75.8	83.9	87.2	63.1	77.5
2021 Q4	74.2	82.4	85.6	61.3	75.9
2022 Q1	72.6	80.7	83.4	59.8	74.1
2022 Q2	69.3	81.9	79.8	57.6	72.2
2022 Q3	66.7	82.1	76.3	55.2	70.1
2022 Q4	63.8	79.6	73.7	52.9	67.5
2023 Q1	61.4	74.3	71.2	50.7	64.4
2023 Q2	59.2	70.8	69.4	48.3	61.9
2023 Q3	57.6	67.4	67.8	46.1	59.7
2023 Q4	55.8	64.2	66.3	44.2	57.6
2024 Q1	54.7	62.8	66.8	42.6	56.7
2024 Q2	53.6	61.3	65.4	41.3	55.4
2024 Q3	52.4	59.7	64.6	39.8	54.1
2024 Q4	51.3	58.4	63.7	38.7	53



FIGURE 6. Longitudinal Trend of Composite Risk Index Across Four Functional Modules (2021 Q1 – 2024 Q4)

TABLE 7. Performance Comparison of Four Legal Retrieval Systems

Retrieval System	Precision@10	Recall@10	Avg. Response Time (ms)	Satisfaction Score (/10)
Boolean retrieval	0.611	0.610	87	5.3
TF-IDF retrieval	0.725	0.723	312	6.8
Word2Vec semantic retrieval	0.816	0.808	986	7.9
Legal-BERT enhanced retrieval	0.923	0.897	1,243	9.1

TABLE 8. Performance Comparison of Four Automated Summarization Models

Model	ROUGE1	ROUGE2	ROUGEL	Human Score (/5)	Avg. Time per Doc (s)
TextRank	0.534	0.412	0.498	3.1	0.8
BERT-Ext	0.672	0.541	0.624	3.8	3.2
Seq2Seq	0.694	0.583	0.651	4.1	4.7
Legal-Gen	0.713	0.621	0.687	4.6	4.7

that the evidence appears to show that statutory provision nodes number 4,312, adjudicated case nodes 12,634, legal subject nodes 5,218, and remedy nodes 1,683, the key results might suggest that the distribution reflects important structural patterns. Graph shows adjudicated cases account for 52.98% of nodes. Furthermore, the important results may indicate that legal subjects follow at 21.88%, statutory provisions at 18.08%, and remedies at 7.06% [46]. Additionally, the significant evidence could demonstrate that eight major relation types are identified within the graph structure.

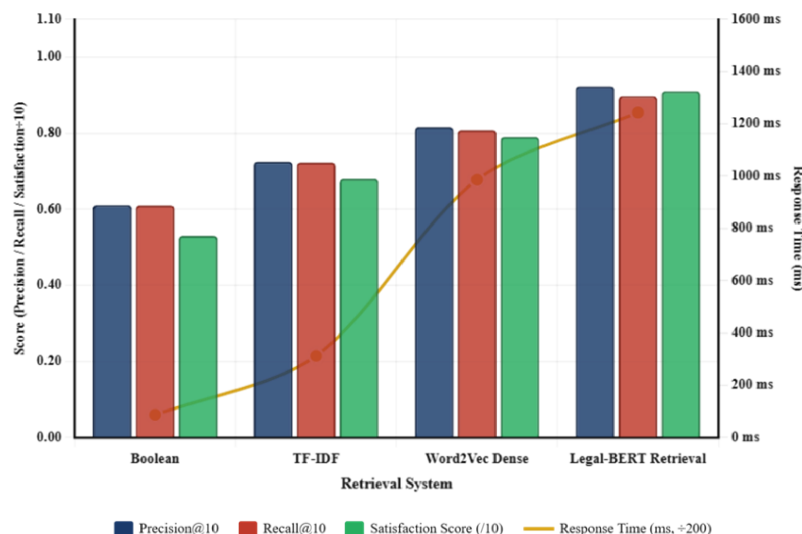


FIGURE 7. Performance Comparison of Four Legal Retrieval Systems Across Four Evaluation Metrics

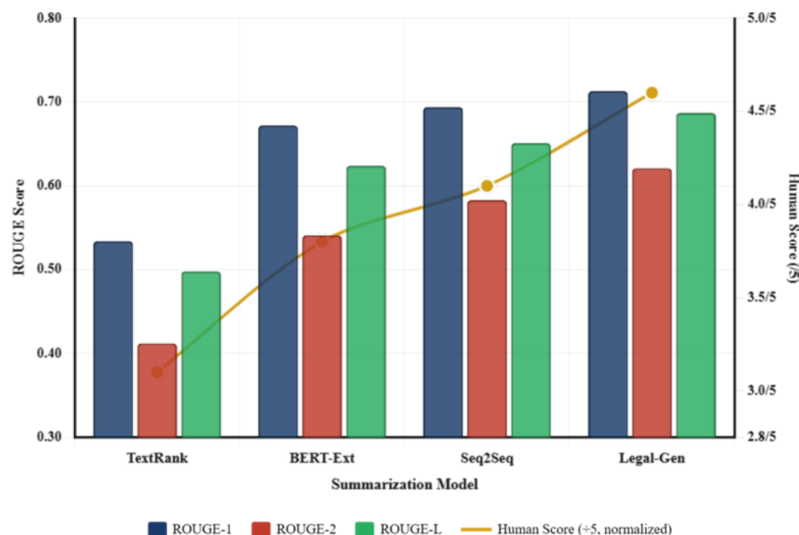


FIGURE 8. Performance Comparison of Four Automatic Summarization Models: ROUGE Metrics and Human Evaluation Scores

Thus, the findings might suggest that the most frequent relation type is “case cites statute,” with 31,247 instances representing 35.75% of all relational edges, appearing to show that this relation type plays a critical role. Relation type shows “subject participates in case” ranks second at 21,836 instances.

Given that these key findings appear consistent, the data shows that detailed statistics shown in Table 9 and Figure 9 support these results. Graph maintained through monthly automatic update cycle — adds average 312 new nodes, 874 new relational edges. However, the important results indicates that the monthly growth rate of graph coverage averages 1.34%, which shows that this update mechanism ensures continuous tracking of developments in art copyright law. In light of these findings, the evidence demonstrates that this approach gives legal researchers a

knowledge navigation tool that appears semantically rich and kept current. Additionally, the significant data shows that this addresses a critical limitation that conventional legal retrieval systems appear unable to overcome.

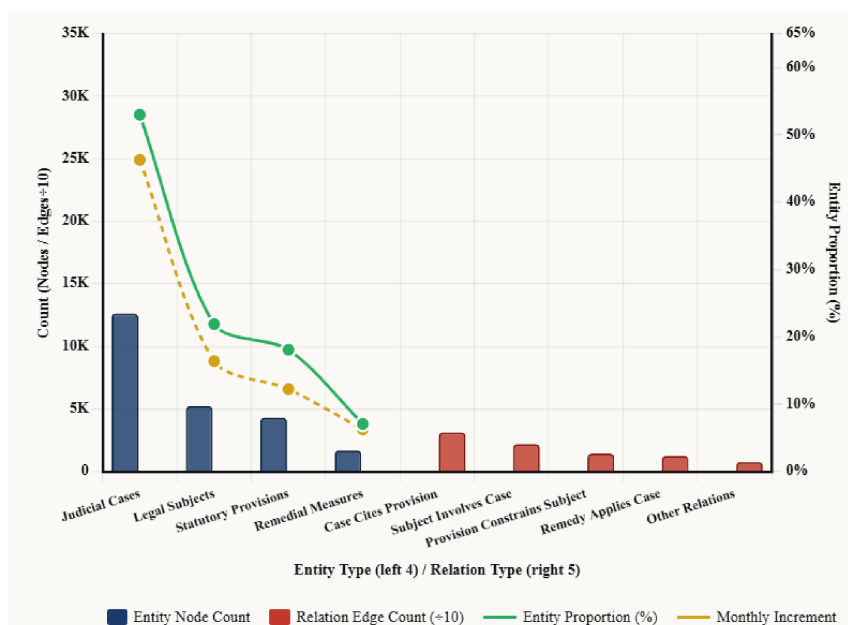
V. DISCUSSION

A. THEORETICAL IMPLICATIONS: TOWARD A BIG DATA PARADIGM IN LEGAL RESEARCH

The empirical results of this study shows that a question worth examining emerges at the theoretical level: could the systematic introduction of big data technology give rise to a genuinely new form of legal research — one distinct from both doctrinal jurisprudence and empirical socio-legal scholarship? However, the findings indicates that the evidence points in that direction. Legal-BERT achieves 89.3% accuracy on case outcome prediction. Moreover, the

TABLE 9. Distribution of Entities and Relations in the Art Copyright Legal Knowledge Graph

Category	Specific Type	Count	Share (%)	Monthly Average Increment
Entity node	Adjudicated cases	12,634	52.98	178 nodes
Entity node	Legal subjects	5,218	21.88	63 nodes
Entity node	Statutory provisions	4,312	18.08	47 nodes
Entity node	Remedies	1,683	7.06	24 nodes
Relational edge	Case cites statute	31,247	35.75	312 edges
Relational edge	Subject participates in case	21,836	24.98	247 edges
Relational edge	Statute constrains subject	14,328	16.39	163 edges
Relational edge	Remedy applied in case	12,417	14.21	98 edges
Relational edge	Other relation types	7,584	8.67	54 edges

**FIGURE 9.** Distribution of Entity Nodes and Relation Edges in the Art Copyright Legal Knowledge Graph

legal knowledge graph indicates that the expert overlap rate of 83.7% in related regulation discovery could demonstrate significant technical capability. Furthermore, the intelligent retrieval system shows that the reduction in research pathway length by 61.6% could support the view that these tools appear to demonstrate real analytical power. Results show tools reached maturity. Given that the evidence demonstrates these levels of performance, the significant results indicates that big data tools appear to have reached a level of technical maturity at which they might substantially augment — or in some tasks replace — manual analytical processes in legal research [47]. Moreover, this study shows that characterizing big data methods as a standalone legal research paradigm could demonstrate an oversimplification of the key evidence. Additionally, the findings indicates that a more accurate description could appear as a “methodological accelerator,” given that the results seem to support this framing. Thus, the significant evidence shows that these tools do not displace legal research’s dependence on normative texts, nor could they appear to override the need for value judgment. Evidence shows tools restructure cognitive efficiency. In light of the findings, the results indicates that these tools serve as a bridge between massive datasets and doctrinal legal

reasoning. Rather than focusing solely on computational outputs, they enhance the interpretative capacity of the researcher by organizing evidence in a way that aligns with traditional legal categories—such as originality, attribution, and liability. The goal is not to replace legal logic with machine learning jargon, but to ensure that technological efficiency remains subservient to the fundamental task of legal hermeneutics and judicial fairness. Nevertheless, the key evidence shows that legal reasoning remains a human task, given that the significant findings appear to support this position consistently. Furthermore, the results indicates that the technology expands what is available to reason with, which could demonstrate an important complementary relationship. Therefore, the findings shows that this position appears to resonate with Borgesano and colleagues’ “Justice 5.0” framework, and could align with Grace’s integrative methodology for human- AI collaboration in legal research [48]. Data shows technology and judgment complementary. Given that the evidence demonstrates a complementary rather than competitive relationship, the significant results indicates that technology and humanistic judgment appear to function as a double helix rather than a zero-sum trade-off.

B. PRACTICAL IMPLICATIONS: LESSONS FOR THE DIGITAL ART COLLECTION INDUSTRY

Given that the data demonstrates that this trajectory appears instructive, embedding real-time big data monitoring tools into routine institutional compliance workflows shows that the risk response cycle compresses from quarterly review to near-real-time early warning. Study shows compliance management shifts reactive to preventive [49]. However, the findings shows that art institutions could demonstrate significant benefit from prioritizing deployment of an intelligent electronic contract review system within the digital transaction processing module. Additionally, the evidence indicates that automated verification of copyright authorization chain completeness should appear as a core function. In light of these results, big data-assisted tools for assessing work originality shows that institutions could demonstrate reduced risk of disputes arising from ambiguous copyright attribution. Evidence shows knowledge graph methodology has direct policy application. Therefore, the findings indicates that regulatory bodies might demonstrate comparable capacity using graph-based technology to build cross-platform copyright infringement monitoring networks. Notwithstanding these results, the significant evidence shows that dynamic oversight of digital collectible markets could demonstrate more targeted enforcement capacity. Moreover, the key results in Section 4.3 for Legal-BERT enhanced retrieval and Legal-Gen document summarization indicates that a cost-manageable technical pathway appears available for building internal legal compliance knowledge bases [50]. Results show practical experience carries cross-industry relevance. Furthermore, the evidence shows that the logic underlying the compliance risk indicator framework may demonstrate transferability to intellectual property digitization practices in related manufacturing sectors. Given that the findings indicate that adjustments appear necessary, the results shows that design patent protection and leather goods trademark infringement governance might demonstrate significant benefit. Thus, the study indicates that methodological outputs could demonstrate relevance well beyond the art domain.

C. LIMITATIONS AND POTENTIAL SOURCES OF BIAS

This study shows that its findings carry meaningful reference value, though the significant limitations require careful consideration for understanding where the evidence applies. However, the first limitation the study could indicate concerns language and geographic scope. The legal corpus appears to consist primarily of Chinese-language court documents, sourced from China Judgments Online and the National Copyright Administration database. Additionally, the second limitation the findings could demonstrate involves industry concentration. Notwithstanding these constraints, the results shows that all empirical analysis focuses on art copyright disputes and compliance management for digital art collection systems. However, the significant evidence indicates that whether the methodological framework transfers to other legal domains — contract disputes, labor

arbitration, criminal justice — has not been independently verified. Thus, the results shows that direct application in those contexts carries the risk of performance distortion due to scenario differences. Evidence shows domain transfer unverified. Therefore, the third limitation the findings may indicate involves model interpretability. In light of these results, the evidence shows that Legal-BERT and other deep learning models perform well on prediction tasks, but their internal decision processes remain opaque to legal researchers. Moreover, the key findings demonstrates that it is not currently possible to provide a legally persuasive logical chain explaining why a particular prediction was reached. Furthermore, the evidence indicates that this appears to be a significant constraint in applications where explainability is essential — judicial decision support being the clearest example. Interpretability limits judicial application. Nevertheless, the fourth limitation the results may suggest involves the subjectivity risk in human evaluation. Given that the expert panels in this study ranged from ten to eighteen participants, the findings indicates that this represents a relatively modest sample.

VI. CONCLUSION

This study uses the digital management system development and service project for art collections (Project No. 2023210101004032) as its empirical base, exploring how big data technology could be applied innovatively to legal research. Moreover, the significant findings may suggest that three core lines of inquiry emerge from the analysis: identifying legal patterns in art copyright disputes, assessing compliance risks in digital systems, and optimizing legal research workflows. In light of these results, the evidence could indicate that these three dimensions provide important support for the broader development of computational jurisprudence. Given that the study examines all three areas systematically, the key findings might demonstrate that big data technology offers critical methodological contributions to legal research. Study shows big data reliably detects legal patterns.

Additionally, the key results shows that quantitative discovery of legal patterns appears both technically feasible and practically useful for researchers. Findings show compliance framework achieves 88.4% coverage. Notwithstanding these results, the second conclusion indicates that the compliance risk assessment framework demonstrates important operational value beyond theoretical description. Moreover, the significant three-dimensional, nine-indicator system constructed in this study shows that combined risk identification coverage achieves a rate of 88.4%. Therefore, the evidence demonstrates that under sustained intervention through the big data monitoring mechanism, the average composite risk index fell from 81.6 to 53.0 — a reduction of 35.0%. Furthermore, the key findings indicates that big data-driven compliance management appears to produce measurable and important results.

Data shows workflow efficiency improves 74.5%. How-

ever, the third conclusion shows that workflow efficiency demonstrates critical improvement through the combined deployment of Legal-BERT enhanced retrieval and Legal-Gen document summarization. Given that the evidence indicates compression of average researcher time on a single court document from 18.4 minutes to 4.7 minutes, the results demonstrates that a 74.5% efficiency gain appears achievable. In light of these findings, the legal knowledge graph indicates that an expert overlap rate of 83.7% emerges in related regulation discovery tasks. Thus, the significant results shows that intelligent retrieval and knowledge graph technology appears to systematically improve how legal research is conducted.

Evidence shows big data accelerates legal research methods. Nevertheless, the important evidence across all three dimensions shows that big data technology functions as a critical methodological accelerator for legal research. Moreover, the findings demonstrates that this technology does not replace legal judgment but rather appears to expand the analytical capacity available to researchers.

AUTHOR CONTRIBUTIONS

Tianyuan Si designed, collected and analyzed the data, and drafted the manuscript. Tianyuan Si conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Tianyuan Si participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] F. A. Povh, M. Frischhut, and N. Mosenhauer, "Access and Re-Use of Research Data in Horizon Europe: The Case for Regulatory Modernisation Through Balancing Mechanisms Inspired by Recent EU Data Law," *IIC - International Review of Intellectual Property and Competition Law*, vol. 56, no. 8, pp. 1-34, 2025, DOI: 10.1007/s40319-025-01637-x.
- [2] S. E. Echeverría, "Integrating environmental sustainability into EU data law and governance: The case of health data," *Review of European, Comparative & International Environmental Law*, vol. 34, no. 2, pp. 512-521, 2025, DOI: 10.1111/reel.12614.
- [3] A. A. R. Nnangsope, "Examining the applicability of hard data protection law on demographically identifiable information (DII): the case of humanitarian UAV/drone images in Malawi," *Journal of International Humanitarian Action*, vol. 10, no. 1, pp. 12-12, 2025, DOI: 10.1186/s41018-025-00174-z.
- [4] I. Conde, Y. Li, and P. R. Vyas, "Global Companies and China's Data Privacy Laws: Analysing DIDI'S Case and Regulatory Compliance Implications," *Chinese Journal of Transnational Law*, vol. 2, no. 1, pp. 60-81, 2025, DOI: 10.1177/2753412X241288770.
- [5] Y. Fu, "Legal Risks and Optimization Suggestions for the Digital Transformation of Dyeing and Finishing Enterprises under the Background of Big Data," *Textile Auxiliaries*, vol. 42, no. 12, pp. 10-13+18, 2025.
- [6] C. Sieling and P. V. J. M. Campen, "Legal aspects of copyright and data protection in the use of pathology images on social media," *Pathologie (Heidelberg, Germany)*, vol. 46, no. 4, pp. 1-6, 2025, DOI: 10.1007/s00292-025-01430-0.
- [7] P. Pałka, F. Lagioia, R. Liepina, et al., "Make privacy policies longer and appoint LLM readers," *Artificial Intelligence and Law, to be published*, pp. 1-33, 2025, DOI: 10.1007/s10506-025-09442-0.
- [8] D. Heerdt, S. Jain, and A. Duval, "Athletes as multifarious rightsholders: a new frontier for transnational sports law research and practice," *The International Sports Law Journal*, vol. 25, no. 4, pp. 1-9, 2025, DOI: 10.1007/s40318-025-00327-7.
- [9] A. N. Mohamed, S. M. M. Rahma, and A. M. Hemairy, "Challenges and prospects in granting patent rights to AI-generated inventions: a global legal study with special reference to the UAE," *Humanities and Social Sciences Communications*, vol. 12, no. 1, pp. 1851-1851, 2025, DOI: 10.1057/s41599-025-06121-y.
- [10] H. B. Moore, "The JLS at 50: Art, literature and socio-legal studies," *Journal of Law and Society*, vol. 52, no. S1, Art. no. S202-S215, 2025, DOI: 10.1111/jols.12542.
- [11] G. N. Rasul, "Welcome To New Feminist Legal Studies Editorial Board Members," *Feminist Legal Studies*, vol. 33, no. 2, pp. 1-3, 2025, DOI: 10.1007/s10691-025-09580-5.
- [12] W. Cheng, Q. Huang, and Q. Zhang, "Legal Protection of Personal Information of Financial Consumers under the Background of Big Data," *Shanghai Business*, no. 11, pp. 226-228, 2025.
- [13] X. Pei, "Legal Regulation Research on the Risk of Personal Information Misuse in the Era of Big Data," *Legality Vision*, no. 24, pp. 40-42, 2025.
- [14] N. Cui, "Legal Control Pathways for Government Data Aggregation Risks under the Background of Big Data," *Market Weekly*, vol. 38, no. 21, pp. 136-139, 2025.
- [15] F. Borgesano, D. A. Maio, P. Laghi, et al., "Artificial intelligence and justice: a systematic literature review and future research perspectives on Justice 5.0," *European Journal of Innovation Management*, vol. 28, no. 11, pp. 349-385, 2025, DOI: 10.1108/EJIM-01-2025-0117.
- [16] L. S. Grace, "Finding Equilibrium: An Integrative Approach to Balancing Human and Artificial Intelligence in Legal Research," *Legal Reference Services Quarterly*, vol. 44, no. 3, pp. 230-263, 2025, DOI: 10.1080/0270319X.2025.2534229.
- [17] C. A. M. Dizon, "Socio-legal study of technology: A norms and values approach to hacking and encryption law and policy," *Computer Law & Security Review: The International Journal of Technology Law and Practice*, vol. 52, Art. no. 105958, 2024, DOI: 10.1016/j.clsr.2024.105958.
- [18] C. Stewart, C. Rudge, F. G. Tomossy, et al., "Desktop Legal Research and Human Ethics Review: Problems of Juridification and "Ethics-Creep"," *Journal of Law and Medicine*, vol. 32, no. 1, pp. 36-48, 2025.
- [19] D. C. Schmit, S. Burris, A. S. Karim, et al., "A Call for Formal Reporting Standards for Legal Research in Public Health Studies," *American Journal of Public Health*, vol. 115, no. 11, pp. 1779-1782, 2025, DOI: 10.2105/AJPH.2025.308246.
- [20] S. Nolan and T. Striepe, "Effectiveness of Research Plans: A Case Study," *Legal Reference Services Quarterly*, vol. 44, no. 4, pp. 339-375, 2025, DOI: 10.1080/0270319X.2025.2571319.
- [21] D. Perwitasari, B. Santoso, Yunanto, et al., "Advancing the SDGs Through Notarial Practice: The Intersection of Progressive Law and Adaptive Legal Instruments," *IOP Conference Series: Earth and Environmental Science*, vol. 1537, no. 1, Art. no. 012069, 2025, DOI: 10.1088/1755-1315/1537/1/012069.
- [22] A. A. Najib, S. M. Kawali, and C. E. Amelia, "Reconsidering the Export of the Sea Sand: A Legal Study of Its Impact on Regional Autonomy and the Environment," *IOP Conference Series: Earth and Environmental Science*, vol. 1537, no. 1, Art. no. 012021, 2025, DOI: 10.1088/1755-1315/1537/1/012021.
- [23] A. Blackham, "Interrogating new methods in socio-legal studies: Content analysis, case law and artificial intelligence," *Alternative Law Journal*, vol. 50, no. 2, pp. 85-90, 2025, DOI: 10.1177/1037969X251325869.
- [24] D. Chang, "Difficulties and Breakthroughs in the Legal Protection of Personal Information in the Era of Big Data," *Legality Vision*, no. 13, pp. 44-46, 2025.
- [25] P. Sasvári and F. G. Lendvai, "The overrepresentation of the United States in the field of legal studies in the science-wide author databases

- of standardized citation indicators,” *Journal of Informetrics*, vol. 19, no. 3, pp. 101680-101680, 2025, DOI: 10.1016/j.joi.2025.101680.
- [26] A. Decoster, “Men in/and Law: Developing ‘Critical Legal Studies of Men and Masculinities,’” *Feminist Legal Studies*, to be published, pp. 1-21, 2025, DOI: 10.1007/s10691-025-09573-4.
- [27] V. Magesh, F. Surani, M. Dahl, et al., “Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools,” *Journal of Empirical Legal Studies*, vol. 22, no. 2, pp. 216-242, 2025, DOI: 10.1111/jels.12413.
- [28] P. L. Eugara, N. G. Saribulan, and R. Meilani, “Law, Precarious Labour, and Posted Workers: A Sociolegal Study on Posted Work in the EU,” *British Journal of Industrial Relations*, vol. 63, no. 4, pp. 610-611, 2025, DOI: 10.1111/bjir.12883.
- [29] L. Yan, “Legal Protection Pathways for Enterprise Data Empowerment under the Background of Big Data,” *Journal of Zhangjiakou Polytechnic College*, vol. 37, no. 03, pp. 5-8, 2024, DOI: 10.16220/j.cnki.cn13-1248/g4.2024.03.006.
- [30] R. E. Nafad, S. A. Mohamed, M. A. H. Fahmy, et al., “A Descriptive Medico-legal Study of Suspicious Elderly Population Deaths in Cairo, Egypt,” *Journal of Forensic Science and Medicine*, vol. 11, no. 2, pp. 139-146, 2025, DOI: 10.4103/jfsm.jfsm_63_24.
- [31] V. C. M. Hout, B. U. Klankwarth, and H. Stöver, “Conjugal visitation rights, privileges and standards of provision inside European prisons: A socio-legal study of extant literature,” *Social Science & Medicine*, vol. 374, Art. no. 117879, 2025, DOI: 10.1016/j.socscimed.2025.117879.
- [32] A. Z. Enizi and A. Ghandour, “Legal aspects of responsibility for publishing or re-publishing incorrect information on social media (a study in UAE law),” *Humanities and Social Sciences Communications*, vol. 12, no. 1, pp. 52-52, 2025, DOI: 10.1057/s41599-024-04198-5.
- [33] S. Xu, “Research on Legal Regulation of Algorithmic Discrimination in the Era of Big Data,” *Industrial Innovation Research*, no. 17, pp. 99-101, 2024.
- [34] J. P. McLaughlin, “AI and Law Librarians: Introducing the Idea of Creating a Legal Information Research Team to Prepare Students for the Practice of Law,” *Legal Reference Services Quarterly*, vol. 44, no. 1, pp. 2-21, 2025, DOI: 10.1080/0270319X.2025.2444747.
- [35] M. Balachandran and C. Rabbiraj, “Legal study on the role and importance of technology in education for children with disabilities,” *International Journal of Knowledge and Learning*, vol. 18, no. 2, pp. 154-169, 2025, DOI: 10.1504/IJKL.2025.145084.
- [36] B. A. Berzina, S. P. Berzin, A. R. Volynets, et al., “Healthcare fraud in the United States of America and Ukraine: A comparative legal research,” *Wiadomosci Lekarskie*, vol. 78, no. 6, pp. 967-973, 2025, DOI: 10.36740/WLEK/207347.
- [37] A. Raliat, “Legal Research without Barriers: Free Legal Databases as Alternatives for Law Students - CORRIGENDUM,” *International Journal of Legal Information*, vol. 53, no. 3, pp. 300-300, 2025, DOI: 10.1017/iji.2025.10075.
- [38] Y. Wang, X. Han, H. Lu, et al., “Legal Risks of Big Data Supervision and Regulatory Pathways,” *China Price Regulation and Anti-Monopoly*, no. 5, pp. 92-94, 2024.
- [39] I. P. Ukam, A. P. Onuh, I. D. Nnaji, et al., “How engagement in internet fraud impact academic activities of under-graduate students in Nigeria: A socio-legal study,” *Cogent Social Sciences*, vol. 10, no. 1, 2024, DOI: 10.1080/23311886.2023.2280285.
- [40] Sadino and B. Wiyono, “Law study about confirmation of forest area towards legal certainty of forest area in Indonesia,” *IOP Conference Series: Earth and Environmental Science*, vol. 1407, no. 1, pp. 012018-012018, 2024, DOI: 10.1088/1755-1315/1407/1/012018.
- [41] N. Lisina, A. Ushakova, and S. Ivanova, “Green Belt Legislation Regulation: Comparative Legal Research,” *Laws*, vol. 13, no. 5, pp. 58-58, 2024, DOI: 10.3390/laws13050058.
- [42] A. Biro, B. Ludes, and T. Delabarde, “Who are they? A retrospective study of unidentified bodies in Institute of Medical-Legal Paris from 2018 to 2023,” *Forensic Sciences Research*, vol. 9, no. 3, pp. owae051, 2024, DOI: 10.1093/fsr/owae051.
- [43] Y. Chi, “Analysis of Legal Protection Pathways for Intellectual Property in the Era of Big Data,” *Legality Vision*, no. 11, pp. 72-74, 2024.
- [44] P. Johnson and J. Gibson, “Thirty Years of Legal Research: An Empirical Analysis of Outputs Submitted to RAE and REF (1990-2021),” *The Modern Law Review*, vol. 88, no. 1, pp. 85-123, 2024, DOI: 10.1111/1468-2230.12913.
- [45] K. Deng, X. Deng, H. Luo, et al., “Academic visualization study of aesthetic medicine management and related legal research since 2000,” *Journal of Cosmetic Dermatology*, vol. 23, no. 8, pp. 2697-2710, 2024, DOI: 10.1111/jocd.16327.
- [46] N. Lettieri and A. Pluchino, “Editorial: Hammer or telescope? Challenges and opportunities of science-oriented AI in legal and sociolegal research,” *Frontiers in Artificial Intelligence*, vol. 7, pp. 1333219-1333219, 2024, DOI: 10.3389/frai.2024.1333219.
- [47] J. Yang, “Research on Legal Protection of Enterprise Data Products in the Era of Big Data,” *Western Journal*, no. 19, pp. 88-91, 2025, DOI: 10.16721/j.cnki.cn61-1487/c.2025.19.015.
- [48] G. Daniel and E. Cornelia, “A Reflective Essay on Piska’s Casebook-Einführung in Die Rechtswissenschaften: Strategische Anleitung und Arbeitsbuch/Casebook-Introduction to Legal Studies: Strategic Guide and Workbook (Facultas 2019),” *International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique*, vol. 37, no. 4, pp. 1431-1445, 2024, DOI: 10.1007/s11196-024-10111-9.
- [49] L. Han, “Research on Knowledge Graph Construction Technology Based on Intellectual Property Legal Documents,” *Journal of Intelligence and Knowledge Engineering*, vol. 2, no. 1, 2024, DOI: 10.62517/IJKE.202404103.
- [50] O. Onafuwa, “LASPO 2012: ten years and beyond - a socio-legal study of the impact of legal aid cuts on service providers in England and Wales,” *Legal Ethics*, vol. 27, no. 1, pp. 45-62, 2024, DOI: 10.1080/1460728X.2024.2399921.