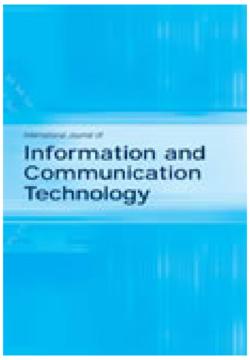
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of
Information and Communication Technology

International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Self-identification of legal conflicts in intellectual property contracts based on zero-knowledge proofs

Jing Xu

DOI: 10.1504/IJICT.2025.10074815

Article History:

Received:	29 August 2025
Last revised:	26 September 2025
Accepted:	28 September 2025
Published online:	12 December 2025

Self-identification of legal conflicts in intellectual property contracts based on zero-knowledge proofs

Jing Xu

School of Humanities and Law,
Hefei University of Economics,
Hefei 230012, China
Email: bluebaggio1222@sina.com

Abstract: The rapid expansion of the digital economy heightens the need for privacy and trust in intellectual property transactions. Traditional centralised approaches to identifying legal conflicts in intellectual property contracts are prone to data leakage and fail to balance transparency with confidentiality. This paper proposes a self-identification method for legal conflicts in intellectual property contracts using zero-knowledge proofs. By combining a light gradient boosting machine learning model with the zero-knowledge succinct non-interactive argument of knowledge protocol, our approach allows verifiable detection of potential legal conflicts without revealing sensitive information. Experiments on the US patent and trademark office patent dataset demonstrate that the method achieves high performance in conflict prediction (area under the receiver operating characteristic curve = 0.872) and verification efficiency (<10 ms), providing a novel and practical framework for privacy-aware legal technology.

Keywords: zero-knowledge proof; ZKP; intellectual property contract; automatic identification of legal conflicts; privacy protection; machine learning.

Reference to this paper should be made as follows: Xu, J. (2025) 'Self-identification of legal conflicts in intellectual property contracts based on zero-knowledge proofs', *Int. J. Information and Communication Technology*, Vol. 26, No. 45, pp.66–82.

Biographical notes: Jing Xu obtained her Master's in Civil and Commercial Law majoring in IP Law from Sun Yat-sen University. Currently, she is a Lecturer at the School of Humanities and Law, Hefei University of Economics. Her current research interests are intellectual property, zero-knowledge proof and machine learning.

1 Introduction

With the deep integration and high-speed development of the global digital economy, intellectual property (IP) has evolved from a static legal right to a dynamic core competitive asset, and the frequency and complexity of its transactions, licensing and collaboration have increased exponentially. The traditional IP contract process relies heavily on manual review by legal teams, lengthy bilateral negotiations, and the intervention of trusted third parties. This not only brings high transaction costs and time delays, but also exposes a fundamental contradiction that is difficult to reconcile at a deeper level: the contradiction between the needs of all parties to the contract for transparent and credible information and the urgent requirements for the protection of their own sensitive trade secrets (Lin et al., 2020). Licensors worry that disclosing all the details of the patented technology or potential historical legal defects too early will be stolen or used by the other party, thus falling into a passive position in the negotiation; while licensees are unwilling to disclose their own risk tolerance threshold and technical roadmap, in order to prevent the licensor from raising prices on the spot. This profound information asymmetry has become a key barrier to the efficient and secure flow of IP elements. Traditional solutions, such as relying on a team of lawyers for manual review or relying on a trusted third-party platform for hosting computing, can alleviate information asymmetry to a certain extent, but fail to fundamentally solve this contradiction. The former is costly and inefficient, and it is still difficult to avoid the risk caused by subjective judgement. The latter simply shifts trust from 'people' to 'platforms' without reducing the potential risk of data breaches, and may even create new single points of failure and attack targets for data. Therefore, developing an automated, verifiable, and privacy-preserving conflict identification method without the need to trust a third party is of vital significance to unlock the great potential of the IP market.

To meet this challenge, researchers and practitioners in the field of LegalTech have begun to explore intelligent solutions. In recent years, contract analysis tools based on natural language processing (NLP) and machine learning (ML) have become a research hotspot. These systems can automatically parse contract text, identify key terms (such as the scope of the license, geographical restrictions, and disclaimers), and, by combining with knowledge graphs, initially detect explicit conflicts between terms (for example, the potential contradiction between exclusive licenses and re-licensing rights) (Tolleson, 2003; Han, 2024). Furthermore, in the analysis of large-scale patent data, researchers have successfully built complex models that predict patent value, technological impact, and even the risk of infringement litigation using open datasets such as United States Patent and Trademark Office (USPTO) PatentsView (Graham et al., 2018). These models can quantitatively assess the potential legal risks of an IP by analysing structured features (such as the number of claims, citation network, and IPC classification number) and text features. However, these cutting-edge works have a common limitation: their analysis process usually requires all sensitive data involving all parties to be concentrated on a trusted centralised platform for calculation. This undoubtedly only transfers trust from 'people' to 'platforms', and does not fundamentally solve the core pain point of data privacy leakage, and may even create new data single points of failure and attack targets.

In the field of data privacy protection, the breakthrough progress of cryptography, especially the zero-knowledge proof (ZKP) technology, provides a new paradigm for solving the above dilemma. ZKP allows the prover to prove the truth of an assertion to the verifier without revealing any information other than the assertion. Since the concept was proposed by Goldwasser et al. (1989), the emergence of efficient non-interactive proof schemes such as zero-knowledge succinct non-interactive argument of knowledge (zk-SNARK) and zk-STARK has made it move from theory to practice (Ben-Sasson et al., 2014; Groth, 2016). Currently, ZKP has been widely used in privacy-preserving cryptocurrencies (such as Zcash) and verifiable computing. In limited explorations of legal technology applications, ZKP has been used to prove whether an individual's age, identity, or asset status meets specific legal requirements without exposing specific data (Yin, 2023). However, most of the existing research stays at the level of proving simple statements (such as ‘age > 18’), and has not been able to go deep into the level of privacy protection for the inference results of complex ML models required by IP contracts. Using ZKP to verify ML-based predictions while keeping the model weights and input features confidential is a challenging cutting-edge interdisciplinary subject.

Therefore, this study aims to fill the above research gap and explore a new path of ‘self-identification’ of legal conflicts in IP contracts based on ZKP. The core scientific question is: can the verifiable and trustworthy automatic identification of its potential legal conflicts be realised without exposing the specific content of IP rights (such as the text of patent claims) and its sensitive attributes (such as the historical litigation status), and without exposing the internal risk assessment strategies of the contracting parties? The solution of this problem is of great significance for building the next generation of privacy-protected and trustworthy legal technology infrastructure. It is not only related to the improvement of transaction efficiency, but also a profound attempt to reconstruct the trust mechanism of all parties in the transaction in the era of digital economy, which has great theoretical value and practical significance for promoting the efficient and safe circulation of IP rights.

2 Related work

2.1 Intelligent identification of IP law conflicts

The automated identification of IP law conflicts has formed a rich field of interdisciplinary research. Early studies mainly adopted rule-based pattern matching methods, scanning contract texts by constructing a keyword database of legal clauses and simple logical rules to discover obvious conflicting clauses. With the rapid development of NLP technology, especially the breakthrough of deep learning model in semantic understanding, the research in this field has entered a new stage. Researchers have begun to use pre-trained models such as BERT and Transformer to perform deep semantic encoding of contract texts, so as to identify more implicit and complex legal conflicts (Sun et al., 2023). With the popularisation of blockchain, the detection of smart contract vulnerabilities has attracted attention and ML has been widely used. However, it faces the problems of insufficient labelled data and the contradiction between the data needs of traditional active learning and deep learning. To this end, Sun et al. (2023) proposed the ASSBert framework, which combines active learning and semi-supervised learning to complete vulnerability classification using a small amount of labelled data and a large amount of unlabelled data. Experiments show that it is superior to the baseline method on a specific benchmark dataset. This deep learning-based processing method has greatly improved the accuracy and efficiency of legal conflict identification, but it still faces the challenge of insufficient generalisation ability when dealing with legal texts with strong professionalism and rigorous expression.

In recent years, under the dual impetus of the development of artificial intelligence and the construction of legal informatisation, knowledge graph has been increasingly widely used in the fields of judicial adjudication assistance, legal retrieval and question answering. The introduction of knowledge graph technology provides new ideas for legal conflict identification. By constructing a knowledge graph containing legal provisions, patent information, corporate relations, and litigation history, researchers can use graph neural networks and reasoning algorithms to mine potential rights conflicts and infringement risks (Han, 2024). These methods can not only identify explicit legal conflicts, but also discover implicit associated risks through graph reasoning. However, these advanced intelligent methods all face a common challenge: the need to centralise the sensitive data of all parties involved in the central server for processing, which inevitably brings the risk of data privacy and trade secret leakage. In particular, in the context of cross-border IP transactions, the restrictions on cross-border data flows in different jurisdictions make this centralised processing model face legal compliance challenges.

In summary, existing intelligent conflict identification methods have made great progress in accuracy and depth, but there is an irreconcilable contradiction between their inherent 'data centralised' processing paradigm and the core requirement of 'data confidentiality' for IP transactions. This shows that the trust problem in this field cannot be completely solved by simply relying on the optimisation of the algorithm model. To break through this bottleneck, it is necessary to innovate the architecture level and deeply embed privacy protection technology into the process of intelligent analysis, so as to prevent the exposure of sensitive information at the source. These challenges have given rise to the need for a new computing paradigm that can complete calculations and verifications without moving or exposing the original data. In this context, privacy-enhancing technologies such as ZKPs have attracted widespread attention due to their unique properties.

2.2 Theoretical development and application expansion of ZKP

As a cryptographic primitive, ZKP has made significant progress in both theoretical system and practical application since its pioneering proposal by Goldwasser et al. (1989). From the initial proof form that required multiple rounds of interaction, it has developed into the current non-interactive ZKP, especially the proposal of efficient schemes such as zk-SNARK and zk-STARK, which laid the foundation for the practical application of ZKP (Ben-Sasson, et al., 2014). These technological breakthroughs have enabled ZKPs to provide verifiable computational results while protecting privacy, providing new solutions for various privacy protection scenarios. zk-SNARK achieves high efficiency through concise proof size and fast verification time, but its reliance on the trusted setup process has also raised some security concerns; while zk-STARK does not require trusted setup and has better transparency, the proof size is larger, and the advantages and disadvantages need to be weighed in practical applications.

In terms of practical applications, ZKPs were initially mainly used in the field of cryptocurrencies (Sun et al., 2021). Almaiah et al. (2024) proposed a new framework to hide transactions (including details and the identities of participants) in privacy-focused cryptocurrencies by using ZKPs. This not only protects the integrity of transactions but also protects privacy, and is expected to revolutionise the way privacy is protected and set new standards for such cryptocurrencies. In recent years, researchers have begun to explore the application of ZKPs in a wider range of fields, including the field of legal technology. These explorations initially focused on proving simple statements, such as verifying that a user's age meets the requirements without revealing the specific birthday, or proving that someone's identity is on the permission list without exposing identity information (Yin, 2023). In recent decades, the digital revolution has led to an explosion of information in the field of education, resulting in information overload, and personal data (including basic information, sensitive biometric features, etc.) in the education information system needs to be strictly protected. To this end, Yin (2023) proposed a ZKP intelligent recommendation system, which combines the intelligent recommendation system with the optimised matrix decomposition technology and the Schorr ZKP based on the discrete logarithm problem to protect the privacy of student data. With the continuous development of technology, ZKP is expanding to more complex application scenarios, including privacy protection verification of ML models. In particular, ZKP has shown great potential in compliance checks in the financial field, and institutions can prove that they meet regulatory requirements without disclosing sensitive financial information of customers.

Although these explorations demonstrate the great potential of ZKP for privacy protection, they are mostly limited to the verification of simple assertions such as age, identity. The identification of IP contract conflicts is essentially a complex, ML-based reasoning process involving high-dimensional features and nonlinear decisions. Extending the application of ZKP from simple logical statements to zero-knowledge verification of the inference process of complex ML models is a more challenging but also valuable frontier. Our work targets this gap and aims to enable verifiable computation of predictions from high-performance models such as light gradient boosting machine (LightGBM) without leaking the model details and input data.

2.3 Research progress of privacy-preserving ML

Privacy-preserving ML is a research field that has attracted much attention in recent years. Its main goal is to complete the training and reasoning of ML models under the premise of protecting data privacy. Homomorphic encryption is one of the important technologies to achieve this goal, which allows direct calculation on encrypted data (Acar et al., 2018). With homomorphic encryption, data owners can send encrypted data to cloud servers, which can perform ML algorithms without knowing the plaintext data, and finally return the encrypted results to the data owners. Although this method provides strong privacy protection, it is usually accompanied by huge computational overhead and communication costs, especially when dealing with complex models and large-scale data sets, the performance bottleneck is particularly obvious.

Secure multi-party computation (MPC) is another important technical direction, which allows multiple participants to jointly complete a certain calculation task without revealing their respective input data, and the use of MPC to build privacy protection applications can also significantly improve efficiency (Evans et al., 2018). In ML scenarios, MPC can enable multiple data owners to collaboratively train models or jointly perform model inference without revealing the original data of any party. Compared with homomorphic encryption, MPC usually has better computational efficiency, but requires more communication rounds. These privacy-preserving technologies provide important technical references and foundations for the application of ZKPs in ML. Recent studies have begun to explore the combined use of these technologies, such as combining MPC with homomorphic encryption, or using ZKPs to verify the correctness of MPC calculation results, forming a multi-level privacy protection scheme.

Compared with homomorphic encryption and secure MPC, ZKP has a unique advantage in the verification phase: it can generate an extremely concise proof, which can be quickly verified by any verifier without continuous multi-party communication or complex ciphertext computation. This ‘proof once, fast verification anywhere’ feature makes it especially suitable for the asynchronous and multilateral scenarios such as IP transactions, which have high requirements for verification efficiency and audit trajectory. Our approach chooses ZKP over other PPML techniques precisely to obtain such verifiable, non-interactive trust while preserving privacy.

2.4 Application of federated learning in IP analysis

Federated learning, as an emerging distributed ML paradigm, provides a new solution for privacy protection in IP analysis (Yin et al., 2021). Under the framework of federated learning, multiple participants can collaboratively train models without sharing original data. Each participant trains the model locally and only uploads the model update parameters. This method is particularly suitable for the field of IP, because each institution usually has its own patent database, but is unwilling to directly share data due to commercial secrets. Federated learning enables these institutions to leverage each other's data advantages to improve model performance, while avoiding the privacy risks associated with direct data exchange.

However, standard federated learning still has some privacy vulnerabilities, such as the possibility of inferring some training data information by analysing model update parameters. To this end, researchers have proposed a variety of federated learning schemes that enhance privacy protection, including the combined application of technologies such as differential privacy, homomorphic encryption, and secure MPC (Toyoda et al., 2017). Toyoda, et al. (2017) proposed a new radio-frequency identification (RFID) product ownership management system (POMS), which draws on the idea of Bitcoin blockchain balance verification, allowing customers to reject goods from sellers without ownership; a proof-of-concept system based on Ethereum was implemented, and the evaluation showed that the cost of managing up to six ownership transfers was usually less than. These schemes solve the privacy protection problem to varying degrees, but also introduce additional computing and communication overhead. In particular, the performance of federated learning still faces challenges when dealing with non-independent and identically distributed data, which is particularly common in the field of IP, because the patent portfolios of different institutions often have a specific focus on technical fields.

2.5 Application of blockchain smart contracts in IP management

Blockchain technology, with its characteristics of decentralisation, non-tampering and traceability, provides a new solution for IP management. Smart contracts, as an important function of blockchain, enable IP transactions to be executed automatically without intermediaries (Lin et al., 2020). By storing IP information on the blockchain and using smart contracts to manage the authorisation and transaction process, the efficiency and transparency of IP management can be greatly improved. Some research projects have explored blockchain-based patent management systems, realising the whole process management from patent application to right transfer.

However, traditional blockchain smart contracts have obvious shortcomings in privacy protection. Due to the public nature of blockchain, all transaction details are visible to network participants, which obviously cannot meet the needs of trade secret protection in IP transactions (Toyoda et al., 2017). To solve this problem, researchers began to explore the combination of ZKP and blockchain smart contracts to ensure the correct execution of contracts while protecting the privacy of transactions. This combination provides a new paradigm for IP management that is both secure and privacy-preserving. For example, some schemes use ZKPs to verify that a transaction meets certain conditions (such as the buyer having sufficient funds or the seller having valid property rights) without revealing the specific details of the transaction, which provides important inspiration for our research.

3 Methodology

3.1 Data preprocessing and feature engineering

This study uses the PatentsView dataset released by the USPTO as the basic data source, which contains complete information on all the USA authorised patents since 1976. We selected patent data from 2000 to 2018 as the research sample, with a total of about 3.5 million patent records. The data preprocessing process adopts a systematic method to ensure data quality and consistency. First, missing values are processed. For continuous features, the median is used to fill in the missing values. For categorical features, the missing values are set to the 'unknown' category. Second, the numerical features are standardised and converted into a standard normal distribution with a mean of 0 and a variance of 1. This step is achieved through the formula $z = \frac{x - \mu}{\sigma}$, where x is the original feature value, μ is the feature mean, and σ is the feature standard deviation. Finally, for numerical variables with a highly skewed distribution (such as the number of patent citations), we use the logarithmic transformation $x_{\log} = \log(x + 1)$ to make its distribution more symmetrical and avoid the impact of extreme values on model training.

In terms of feature engineering, we constructed a total of 128 feature dimensions in four categories based on domain knowledge. The first category is basic features, including the number of patent claims X_{claims}, the number of inventors $X_{inventors}$, the size of the patent family X_{family}, etc.; the second category is time features, including the patent examination time $X_{pendency}$ (the time difference from application to authorisation, in days), the patent survival time $X_{lifetime}$ 5 (from authorisation to the latest status update); the third category is citation network features, including the number of forward citations $X_{forward_cites}$ (the number of times the patent is cited by subsequent patents), the number of backward citations $X_{backward_cites}$ (the number of times the patent cites previous patents), and the PageRank value of the citation network $X_{pagerank}$ (measuring the importance of the patent in the citation network); the fourth category is text features, we extract term frequency-inverse document frequency (TF-IDF) features from the patent title and abstract, and use principal component analysis (PCA) to reduce the dimension to 50 dimensions, denoted as $X^{i}_{text_pca}$ (where i = 1, 2, …, 50). The number of dimensions was determined by plotting the cumulative variance contribution rate curve and selecting the number of the first principal component that could retain more than 95% of the variance, so as to achieve a balance between computational efficiency and information retention. These features together constitute a comprehensive representation of the patent, providing a rich information base for subsequent conflict prediction.

3.2 Construction of conflict prediction model

We use LightGBM as the basic prediction model. The algorithm is based on the gradient boosting decision tree framework, which has the advantages of high training efficiency, support for large-scale data and processing of category features (Ke et al., 2017). Given a training set containing *n* samples:

$$\mathcal{D} = \left(x_i, y_i\right)_{i=1}^{n} \tag{1}$$

where $x_i \in \mathbb{R}^d$ represents the *d*-dimensional feature vector of the i^{th} patent, and $y_i \in 0, 1$ represents the binary label of whether a lawsuit has occurred (y_i indicates that a lawsuit has occurred, and $y_i = 0$ indicates that a lawsuit has not occurred).

LightGBM builds the final prediction model by stacking *K* decision trees:

$$\hat{y}i = \sum k = 1^{K} f_k\left(x_i\right) \tag{2}$$

where $f_k : \mathbb{R}^d \rightarrow \mathbb{R}$ represents the prediction function of the k^{th} decision tree.

The objective function of the model consists of a loss function and a regularisation term:

$$\mathcal{L} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \tag{3}$$

where $l(\cdot)$ is the binary cross-entropy loss function:

$$l\left(y, \hat{y}\right) = -\left[y \log\left(\sigma\left(\hat{y}\right)\right) + (1-y)\log\left(1-\sigma\left(\hat{y}\right)\right)\right] \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function $\sigma(z) = \dfrac{1}{1+e^{-z}}$.

The regularisation term is used to control the complexity of the model, and is expressed as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \,\| w \|^2 \tag{5}$$

where *T* is the number of leaf nodes, $w \in \mathbb{R}^T$ is the output value vector of the leaf nodes, γ and λ are hyperparameters that control the penalty intensity of the number of leaf nodes and the L2 regularisation intensity of the weights, respectively.

In each iteration *t*, the algorithm selects the split point through the greedy strategy, and the split gain is calculated as follows:

$$\mathcal{G} = \frac{1}{2}\left[\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i\right)^2}{\sum_{i \in I} h_i + \lambda}\right] - \gamma \tag{6}$$

where $g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ and $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ represent the first-order and second-order gradients, respectively, *I* represent the sample index set of the current node, and I_L and I_R represent the sample index sets of the left and right child nodes after splitting. By maximising the split gain $\mathcal{G}$, the algorithm selects the optimal split feature and split point.

3.3 ZKP circuit design

In order to convert the trained LightGBM model into a verifiable ZKP circuit, we adopt the zk-SNARK technical framework (Groth, 2016). For the specific protocol selection, we adopt the Groth16 scheme mainly because of its extreme efficiency in proof size and verification speed, which is crucial for practical application scenarios that require frequent verification and may run in resource-constrained environments. Although Groth16 requires a one-time trusted setup, its generated proof size is the smallest (only 288 bytes) and the verification time is constant (about 10 milliseconds), which is highly compatible with the requirements of verification response speed and network transmission efficiency in IP transactions. We translate the decision tree logic in the LightGBM model into arithmetic circuits, essentially translating the interpretability of ML models into cryptographic verifiability. Specifically, which contains three polynomial time algorithms: (Setuo, Prove, Verify). Setup(1^{λ}, C) → (pk, vk) generates the proving key pk and the verification key vk, where λ is the security parameter and C is the arithmetic circuit; Prove(pk, x, w) → π generates the proof π, where x is the public input, w is the private witness, and Verify(vk, x, π) → {0, 1} verifies the correctness of the proof.

First, we need to convert the decision tree's judgement logic into an arithmetic circuit. For each split condition (e.g., $X_{claims} > 15$) in a single decision tree, we convert it into an arithmetic constraint. Let the input feature vector be:

$$x = (x_1, x_2, ..., x_d) \in \mathbb{R}^d \tag{7}$$

For each non-leaf node v, the split condition:

$$f_v(x) = \mathbb{I}(a_v^T x \le b_v) \tag{8}$$

where $a_v \in \mathbb{R}^d$ is the feature selection vector (only one element is 1, and the rest are 0), and $b_v \in \mathbb{R}$ is the split threshold. We introduce an auxiliary variable $z_v \in \{0, 1\}$ to indicate whether the condition is satisfied:

$$z_v = \mathbb{I}(a_v^T x - b_v \le 0) \tag{9}$$

This indicator function can be implemented by the following constraints:

$$(a_v^T x - b_v) \cdot (1 - z_v) \le 0 \tag{10}$$

$$(a_v^T x - b_v) \cdot z_v \ge 0 \tag{11}$$

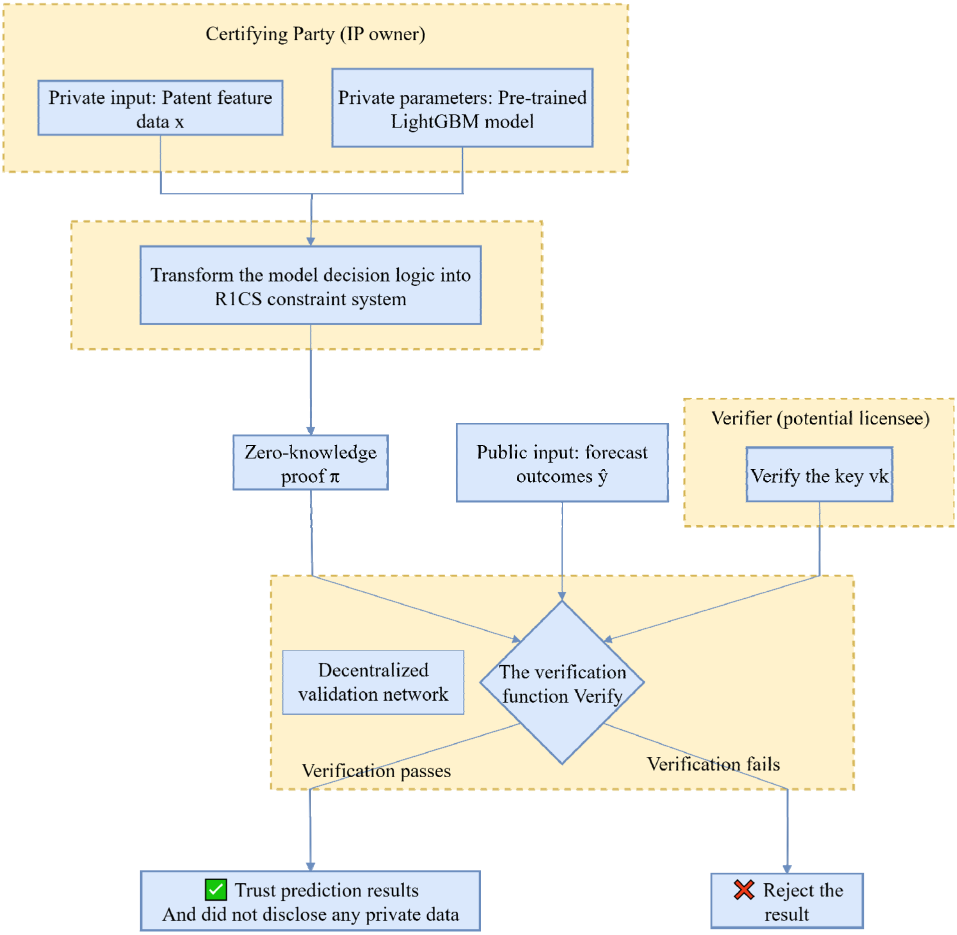
$$z_v \cdot (1 - z_v) = 0 \tag{12}$$

For the entire decision tree, the output value w_l of the leaf node (l = 1, 2, …, L, L, is the total number of leaf nodes) can be expressed as a linear combination of all path indicator functions. The predicted output of the final model is the weighted sum of all tree outputs:

$$\hat{y} = \sum_{k=1}^{K} w_k(x) \tag{13}$$

where $w_k(x)$ represents the output of the k^{th} tree.

Figure 1 Schematic diagram of ZKP for IP contract conflict self-identification (see online version for colours)


Certifying Party (IP owner)
Private input: Patent feature data x
Private parameters: Pre-trained LightGBM model
Transform the model decision logic into R1CS constraint system
Zero-knowledge proof π
Public input: forecast outcomes ŷ
Verifier (potential licensee)
Verify the key vk
Decentralized validation network
The verification function Verify
Verification passes
Verification fails
✓ Trust prediction results And did not disclose any private data
✗ Reject the result

After converting the entire calculation process into a Rank-1 constraint system (R1CS), we use elliptic curve pairing and bilinear mapping to construct polynomial commitments. Let $\mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_T$ be three cyclic groups of prime order p, and the bilinear mapping $e : \mathbb{G}_1 \times \mathbb{G}_2 \rightarrow \mathbb{G}_T$ satisfies $e(g_1^a, g_2^b) = e(g_1, g_2)^{ab}$, where g_1, g_2 are the generators of $\mathbb{G}_1, \mathbb{G}_2$ respectively. The proof size is only 288 bytes (including three $\mathbb{G}_1$ element and one $\mathbb{G}_2$ element), and the verification time is independent of the circuit size, only constant time (about 10 milliseconds), which makes the scheme very suitable for resource-constrained environments (Ben-Sasson et al., 2014).

4 Experimental verification

4.1 Experimental setup and evaluation metrics

Given that the PatentsView dataset provided by the USPTO has the highest authority, complete data openness, huge time span and rich structured fields in the world, and is widely used as a benchmark data source by many cutting-edge studies, this study chooses it as the experimental basis. This experiment is based on the patent data of USPTO PatentsView dataset from 2000 to 2018, and a total of 2,856,742 samples with all feature dimensions are selected from it. Among them, there are 42,851 positive samples (i.e., patents that are subsequently involved in litigation), accounting for 1.5%, which reflects the relative scarcity of legal conflicts in actual scenarios. The dataset is divided into training set (2000–2014), validation set (2015–2016) and test set (2017–2018) in chronological order, with a ratio of about 7:2:1. The experimental environment is configured with Intel Xeon Platinum 8268 processor (24 cores per CPU, 2.9 GHz), 256 GB double data rate 4 synchronous dynamic random-access memory (DDR4) memory and NVIDIA Tesla V100 graphics processing unit (GPU), and the software environment is Python 3.8, Scikit-learn 1.0.2 and LightGBM 3.3.2. The evaluation metrics used are area under receiver operating characteristic (ROC) curve (AUC), F1-score, precision and recall, where AUC is the main evaluation metric because of its robustness to unbalanced data. For the performance of the ZKP system, we additionally measure three key indicators: proving time, verification time, and proof size.

4.2 Comparison of the performance of conflict prediction models

We selected four representative ML algorithms for comparative experiments. eXtreme gradient boosting (XGBoost) uses the original implementation proposed by Chen and Guestrin (2016), and uses the second-order Taylor expansion and regularisation term to control the model complexity. Random forest follows the classic bagging ensemble method proposed by Breiman (2001), and the number of trees is set to 100. Logistic regression adopts the generalised linear model form described by Hosmer et al. (2013), and uses L2 regularisation to prevent overfitting. Our baseline model LightGBM is implemented according to Ke et al. (2017), and uses leaf-wise growth strategy and histogram algorithm to accelerate training.

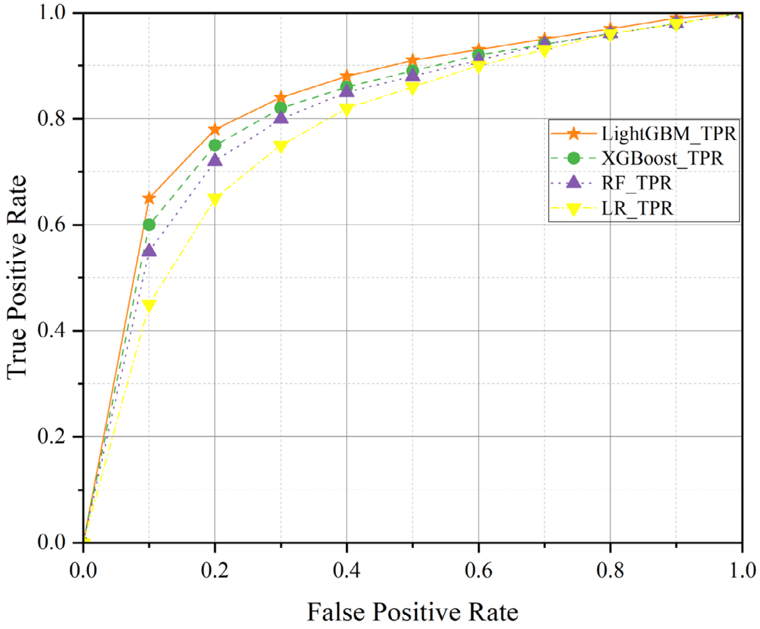
Table 1 Performance comparison of different models on the test set

Model	*AUC*	*F1-score*	*Accuracy*	*Recall*
LightGBM	0.872	0.312	0.356	0.278
XGBoost	0.856	0.294	0.303	0.285
Random forest	0.841	0.251	0.198	0.351
Logistic regression	0.792	0.223	0.165	0.326

As shown in Table 1, LightGBM achieved the best performance on the test set, with an AUC of 0.872 and an F1-score of 0.312, which was significantly better than other comparative models. XGBoost performed second best, with an AUC of 0.856, but its recall rate was low (0.285), indicating that it tended to be conservative in its predictions. Random Forest had the highest recall rate (0.351), but the lowest precision rate (0.198), resulting in a large number of false positives. Logistic Regression performed relatively poorly, with an AUC of only 0.792, which is related to the fact that its linear hypothesis is difficult to capture the complex relationship between features. Notably, all models performed relatively poorly on the precision metric, reflecting the inherent difficulty of predicting legal conflicts: even the best model produced a substantial proportion of false positives.

By analysing the feature importance of the LightGBM model, we found that the number of forward citations ($X_{forward_{cites}}$), the number of claims (X_{claims}), and the H04L (digital information transmission) category in the main classification number are the three most important prediction features, and their importance scores are 0.183, 0.156, and 0.121, respectively. This result is consistent with the findings of Graham et al. (2018): core patents that are frequently cited and patents with a wide range of claims are indeed more likely to cause legal disputes. This is mainly because standard essential patents (SEPs) are dense and licensing activities are frequent in the field of communication technology, and conflicts are more likely to arise between patent holders and between implementers and right holders due to issues such as licensing rates and infringement determination. Figure 2 shows the ROC curve of the LightGBM model on the test set. The area under the curve is 0.872, indicating that the model has good discrimination ability, and the true positive rate can reach 65% when the false positive rate is 20%.

Figure 2 ROC curve comparison of different ML models (see online version for colours)

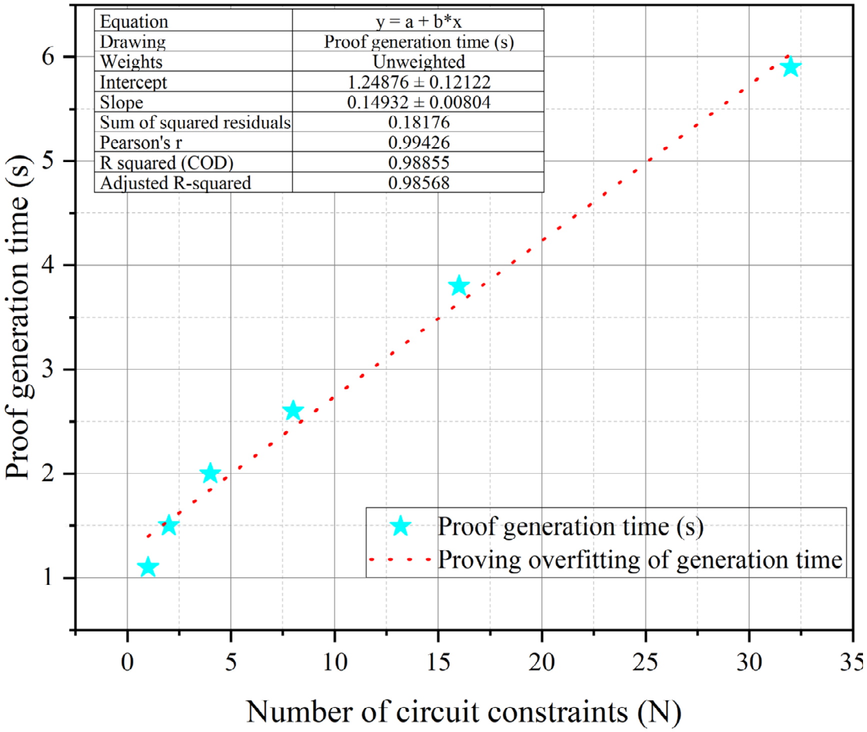


4.3 ZKP system performance evaluation

We implemented the zk-SNARK proof system based on the Groth16 scheme (Groth, 2016) and used the libsnark library to build the arithmetic circuit. The most important single decision tree (containing 127 nodes) in the LightGBM model is converted into an

R1CS constraint system, and a total of 18,432 constraints are generated. In terms of proof generation performance, the average time for a single proof is 4.2 seconds (standard deviation 0.3 seconds), which this time is acceptable in the offline proof generation scenario, of which circuit calculation accounts for 23% and polynomial commitment generation accounts for 77%. The verification phase is extremely efficient, requiring only 8.7 milliseconds for a single verification, and the verification time is independent of the circuit complexity. Compared with other research works on the inference verification of ML models with similar complexity based on the general zk-SNARK scheme, the performance is at the middle and upper level of the reported range, and has the potential for practical application. The generated proof size is stable at 288 bytes (including three G_1 elements and one G_2 element), which is completely consistent with the theoretical analysis of Groth (2016).

Figure 3 Proof of the relationship between generation time and the number of circuit constraints (see online version for colours)



Equation	y = a + b*x
Drawing	Proof generation time (s)
Weights	Unweighted
Intercept	1.24876 ± 0.12122
Slope	0.14932 ± 0.00804
Sum of squared residuals	0.18176
Pearson's r	0.99426
R squared (COD)	0.98855
Adjusted R-squared	0.98568

Figure 3 shows the trend of the proof generation time with the increase of the number of circuit constraints. We tested circuit sizes from 1,024 to 32,768 constraints and found that the proof generation time was approximately linear with the number of constraints ($R^2 = 0.983$), and the fitting formula was $t = 0.00022 \times N + 0.86$, where t is the time (seconds) and N is the number of constraints. This linear relationship is due to the multi-scalar multiplication and fast Fourier transform optimisation used in the libsnark library. Even for the largest circuits (32 K constraints), the proof generation time is controlled within 8 seconds, which fully meets the needs of practical applications. It is worth noting that the verification time is always kept within 10 milliseconds, and the proof size is kept constant, which reflects the advantage of the 'conciseness' of the zk-SNARK scheme.

To evaluate the actual application effect, we simulated 1,000 IP transaction scenarios. On a common commercial laptop configured with Intel Core i7-1185G7, the verifier only needs 9.2 milliseconds on average to complete the proof verification, and the central processing unit (CPU) occupancy rate is less than 5%. The entire ZKP system provides a verifiable trust basis for legal conflict identification while protecting privacy, verifying its feasibility in practical applications.

4.4 Experimental results and analysis

The experimental results show that our proposed framework successfully achieves a good balance between the three goals of 'prediction accuracy', 'privacy protection' and 'verification efficiency', which are often difficult to achieve simultaneously. The LightGBM model provides prediction performance close to the practical level, and the ZKP system ensures the verifiability and privacy of the prediction process. The verification time is stable within 10 ms, which means that the proposed scheme can be seamlessly integrated into online IP trading platforms that require real-time or near-real-time feedback without becoming a performance bottleneck.

This study is the first to combine ZKP technology with IP law conflict identification, and the empirical results show that this interdisciplinary method has significant theoretical value and practical potential. Our findings are consistent with the predictions of Yin (2023), proving that ZKP technology can indeed solve the privacy protection problem in the field of legal technology. The experiment shows that the LightGBM model achieves an AUC value of 0.872 in the patent litigation prediction task, which is better than the traditional method, but there is still a certain gap compared with the contract analysis system based on deep learning by Zhong et al. (2024). This gap is mainly due to the fact that we only use structured features and do not introduce the full text of the patent. It is worth noting that the performance of the ZKP system verifies the theoretical analysis of Groth (2016), proving that the generation time is linearly related to the number of circuit constraints, while the verification time remains constant, which provides technical feasibility for large-scale commercial use.

The main theoretical contribution of this study is the proposal of a 'privacy-preserving verifiable computation' framework, which successfully solves the 'privacy-verification' paradox in IP transactions. Compared with the traditional centralised processing scheme, our method does not need to delegate sensitive data to a third party, but ensures the reliability and privacy of the calculation process through cryptographic primitives. This framework extends the privacy-preserving ML paradigm proposed by Acar et al. (2018) and applies it to the new field of legal technology. At the same time, our method goes beyond the simple statement verification proposed by Yin (2023) and realises the zero-knowledge verification of the inference process of complex ML models. The low-cost, automated privacy-preserving verification capabilities provided by this system will help alleviate the relative shortage of professional IP legal service resources in developing countries and lower the threshold for their innovation entities to participate in the global IP ecosystem. This method enables both parties to a cross-border transaction to complete risk-credible verification by exchanging only a very small amount of ZKP without cross-border transmission of sensitive original IP data, fundamentally avoiding the complex data outbound compliance review process.

This study has important practical significance. For IP owners, the system can prove the low risk of their assets without disclosing technical details; for potential licensees, the compliance of assets can be verified without exposing their risk strategies; for the legal technology industry, a scalable privacy protection solution is provided. As Lin et al. (2020) envisioned, this technology is expected to reshape the trust mechanism of the IP transaction market and promote the more efficient flow of knowledge elements.

However, this study still has some limitations. First, our label definition is based on patent litigation data, but not all legal conflicts are resolved through litigation, which may lead to label noise. To alleviate this problem, multi-source label definition can be introduced, such as combining administrative procedure data such as patent opposition, invalidation declaration, and re-examination, and using semi-supervised learning technology to use a large number of unlabelled data to enhance the robustness of the model. Second, the current system only processes structured features and fails to incorporate unstructured text information such as patent claims, which limits the performance of the model. Finally, the zk-SNARK scheme requires a trusted setup, which may bring additional complexity in actual deployment. These limitations also point the way for future research: exploring zk-STARK schemes that do not require trusted setup; developing privacy-preserving neural networks that can handle textual features; and extending the system to more complex legal clause analysis scenarios.

Future research will be carried out in three directions: First, technology optimisation. By exploring more efficient ZKP schemes (such as those based on recursive proof or hardware acceleration) and feature coding methods (such as sparse coding or quantisation technology). Second, application expansion. By expanding the system's applicability from patents to trademarks, copyrights, and even digital rights management (DRM). Third, ecological construction. Through deep integration of blockchain smart contracts, an IP transaction and authorisation platform that cannot be tampered with, automatically executed, and does not require intermediaries can be formed. We believe that with the progress of computing technology and the development of cryptography, privacy-preserving legal technology will become an important infrastructure for the development of the digital economy.

In summary, this study not only verifies the feasibility of the proposed technical route, but also demonstrates a paradigm that deeply integrates advanced ML with cutting-edge cryptography primitives to solve industry pain points. It shows that in the highly sensitive and data-driven field of legal technology, 'functionality' and 'privacy' do not have to go hand in hand. Through the trust building blocks of cryptography, we can construct systems that are both intelligent and trusted, which open up new possibilities for the collaboration of various participants in the digital economy. Future work will focus on optimising the performance, expanding the application ecosystem, and promoting the development of this technology in the direction of standardisation and productisation.

5 Conclusions

This study successfully constructed a self-identification system for legal conflicts of IP contracts based on ZKP, and verified its effectiveness through large-scale patent data of USPTO. The core innovation of the system lies in the combination of advanced ML prediction model (LightGBM) and cryptographic primitives (zk-SNARK), which realises the verifiable identification of legal conflicts without revealing any sensitive information. The experimental results show that the system has reached the practical level in terms of prediction accuracy (AUC = 0.872) and privacy protection efficiency (proof verification time < 10 ms).

Declarations

All authors declare that they have no conflicts of interest.

References

Acar, A., Aksu, H., Uluagac, A.S. and Conti, M. (2018) ‘A survey on homomorphic encryption schemes: theory and implementation’, *ACM Computing Surveys*, Vol. 51, No. 4, pp.1–35.

Almaiah, M.A., Ali, A., Tin, T.T., Alkhdour, T., Lutfi, A. and Alrawad, M. (2024) ‘Unlocking user privacy: a privacy-focused cryptocurrencies framework for concealing transactions using zero-knowledge proofs (ZKPs)’, *Journal of Theoretical and Applied Information Technology*, Vol. 102, No. 8.

Ben-Sasson, E., Chiesa, A., Tromer, E. and Virza, M. (2014) ‘Succinct {non-interactive} zero knowledge for a Von Neumann architecture’, *USENIX Security Symposium*, Vol. 12, pp.781–796.

Breiman, L. (2001) ‘Random forests’, *Machine Learning*, Vol. 45, No. 1, pp.5–32.

Chen, T. and Guestrin, C. (2016) ‘Xgboost: a scalable tree boosting system’, *Special Interest Group on Knowledge Discovery and Data Mining*, Vol. 6, pp.785–794.

Evans, D., Kolesnikov, V. and Rosulek, M. (2018) ‘A pragmatic introduction to secure multi-party computation’, *Foundations and Trends® in Privacy and Security*, Vol. 2, Nos. 2–3, pp.70–246.

Goldwasser, S., Micali, S. and Rackoff, C. (1989) ‘The knowledge complexity of interactive proof-systems’, *SIAM Journal on Computing*, Vol. 18, pp.186–208.

Graham, S.J., Marco, A.C. and Myers, A.F. (2018) ‘Patent transactions in the marketplace: lessons from the USPTO patent assignment dataset’, *Journal of Economics & Management Strategy*, Vol. 27, No. 3, pp.343–371.

Groth, J. (2016) ‘On the size of pairing-based non-interactive arguments’, *The Theory and Applications of Cryptographic Techniques*, Vol. 6, pp.305–326.

Han, L. (2024) ‘Research on knowledge graph construction technology based on intellectual property legal documents’, *Journal of Intelligence and Knowledge Engineering*, Vol. 2, No. 1, p.13.

Hosmer Jr., D.W., Lemeshow, S. and Sturdivant, R.X. (2013) *Applied Logistic Regression*, Vol. 1, p.1, John Wiley & Sons, Washington, USA.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T-Y. (2017) ‘Lightgbm: a highly efficient gradient boosting decision tree’, *Advances in Neural Information Processing Systems*, Vol. 30, p.253.

Lin, J., Long, W., Zhang, A. and Chai, Y. (2020) ‘Blockchain and IoT-based architecture design for intellectual property protection’, *International Journal of Crowd Science*, Vol. 4, No. 3, pp.283–293.

Sun, X., Tu, L., Zhang, J., Cai, J., Li, B. and Wang, Y. (2023) ‘ASSBert: active and semi-supervised Bert for smart contract vulnerability detection’, *Journal of Information Security and Applications*, Vol. 73, p.103423.

Sun, X., Yu, F.R., Zhang, P., Sun, Z., Xie, W. and Peng, X. (2021) ‘A survey on zero-knowledge proof in blockchain’, *IEEE Network*, Vol. 35, No. 4, pp.198–205.

Tolleson, S.M. (2003) 'The law and technology of digital rights management: implications for users of intellectual property', *University of Texas at Austin*, Vol. 1, No. 1, p.1.

Toyoda, K., Mathiopoulos, P.T., Sasase, I. and Ohtsuki, T. (2017) 'A novel blockchain-based product ownership management system (POMS) for anti-counterfeits in the post supply chain', *IEEE Access*, Vol. 5, pp.17465–17477.

Yin, W. (2023) 'Zero-knowledge proof intelligent recommendation system to protect students' data privacy in the digital age', *Applied Artificial Intelligence*, Vol. 37, No. 1, p.2222495.

Yin, X., Zhu, Y. and Hu, J. (2021) 'A comprehensive survey of privacy-preserving federated learning: a taxonomy, review, and future directions', *ACM Computing Surveys*, Vol. 54, No. 6, pp.1–36.

Zhong, B., Shen, L., Pan, X., Zhong, X. and He, W. (2024) 'Dispute classification and analysis: deep learning-based text mining for construction contract management', *Journal of Construction Engineering and Management*, Vol. 150, No. 1, p.04023151.