International Journal for Housing Science and Its Applications
Publish September 11, 2025. Volume 47, Issue 1 Pages 865-877

https://doi.org/10.70517/ijhsa47175

Sentencing standardization based on judicial big data using gradient boosting decision trees

Ying Chieh Lin[1,*] and Shaojun Liu[1]
[1]School of Civil, Commercial, and Economic Law, China University of Political Science and Law
Corresponding authors: (e-mail: inja@asdgac.org).

Abstract In the era of digital justice, the integration of big data analytics into sentencing decisions has emerged as a key direction for enhancing judicial transparency and fairness. This paper proposes a novel sentencing standardization framework based on judicial big data and interpretable machine learning. Focusing on online fraud adjudication documents from the Chinese judiciary, we construct a domain-specific database using a hybrid method of keyword-based pattern matching and association rule analysis to extract structured features such as criminal intent, means, economic loss, and mitigating factors. These features are encoded into machine-readable vectors and fed into a LightGBM-based gradient boosting decision tree (GBDT) model to predict sentencing outcomes. Extensive experiments using real-world fraud cases demonstrate the model's high predictive performance, with R^2 scores reaching 0.98 and minimal average deviation. A series of visual and statistical evaluations—including boxplots, Taylor diagrams, and regression fits—validate the model's robustness and its ability to replicate human sentencing logic.

Index Terms judicial big data, sentencing prediction, LightGBM, legal AI, pattern recognition, Chinese court decisions, online fraud, standardization model

I. Introduction

The impact of the big data wave on the human economy and society is far greater than that of human capital, technology, and other factors. The use of big data technology in a variety of fields has become a national strategy, and the extent of its application is a significant indicator of a nation's overall competitiveness [1]. Big data technology offers a historic chance for judicial modernization in the area of judicial adjudication while fostering social advancement and growth [2]. Thus, the following are crucial topics that require in-depth discussion: what is the current state of big data in judicial adjudication, how to comprehend the importance and value of big data technology in judicial adjudication, how to resolve issues encountered when applying big data technology in judicial to encourage the modernization of judicial adjudication with big data technology.

In recent years, the judicial system has been gradually combined with a variety of modern information technologies. Modern technologies have supported the promotion of close integration of the judicial field with the trend of the times, but at the same time they have also caused negative effects [3], [4]. For example, since justice itself has fixed properties, it is subject to different degrees of impact in the process of integration with more modern inform and non-referential nature of risks, full coverage cannot be achieved between the currently existing legal norms and information technology ethics. In this context, how to recognize the possible potential risks of judicial big data in the application process has become an urgent problem to be solved in the process of realizing the new technology risk prevention [5], [6]. For the special nature of the judicial scene, the use of traditional information technology has been unable to accurately assess the risks and problems that exist, and for this problem, the paper introduces artificial intelligence technology. With capabilities like prediction and help in a variety of sectors, as well as the ability to optimize data utilization, artificial intelligence technology is currently widely utilized in many different fields. Therefore, the risk assessment method of judicial big data application is examined through this technology, integrating the benefits of applying artificial intelligence technology.

The comprehensive use of big data in judicial adjudication is crucial for enhancing judicial authority, quality, and efficiency as well as for supporting trial system reform, judicial information disclosure, and the development of a rule of law in China [7], [8]. However, the judgment is made by judicial adjudicators based on their own understanding of social and economic dynamics, legal provisions and practical experience in adjudication, and their work is value-judged and empirical, which cannot be quantified [9], [10]. In this paper, while analyzing the value and significance of big data application in judicial adjudication, we propose to objectively treat the problems in big data appig data technology really work in judicial adjudication. The essence of big data technology is the technology that can achieve the functions of analysis, learning, identification, and decision making until replacing human beings by simulating the process of human thinking and information processing [11], [12]. As an emerging technology, its development has a huge impact on the free mind, discretion, and causality in judicial adjudication [13], [14]. Through the principle, characteristics and practice status of big data in judicial adjudication, this paper proposes to reasonably position the relationship between big data and judicial adjudication, establish data thinking, and gradually shift from the judgment of causality to correlation, which offers a theoretical framework for using big data technology in the realm of court decision-making.

In this study, we focus on the fraud crime adjudication documents released by the Judgment Document Network, select some telecommunication network fraud crime cases, and use pattern matching and machine deep learning to quickly identify the case information features. After that, a sentencing specification database is established to store the basic information of the offender, crime purpose, crime mode, crime result, aggravating circumstances, sentence reduction, trial result and other case information. Finally, a standarentencing model is established with the extracted case elements as features and the judgment results as prediction targets, and the test data are compared with the actual data to test whether the model results are accurate.

II. Case information feature extraction by pattern matching and association analysis

A schema is a special type of metadata model for database applications. It is a finite structure of related elements, such as views, tables, classes or collections of XML elements and attributes. Schemas are now used to achieve heterogeneity. These elements are important and necessary for interoperability of data and application systems. In the case of sche be mapped to relevant case elements based on relevant keywords. In judgement text extraction, relevant keywords are selected as criteria for classifying judgement texts and sentences, and different combinations of keywords are used to highlight judgement pattern categories, which are then placed at different nodes for data import. The keywords are selected as shown in Table 1 and the data collection and import process is shown in Figure 1.

Table 1: Keyword selection

Sentence paragraph	Key words	Fall into category
Criminal information	Male, female, ethnicity, year, month, day, household registration, domicile...	Criminal identity information
Criminal Facts Paragraph	Cheat, get, cheat, cheat, get...	Criminal purpose
Criminal Facts Paragraph	Network, mobile phone, card, telephone, impersonation, counterfeiting, fabrication...	Criminal means
Criminal Facts Paragraph	Total...RMB, total...RMB, RMB, yuan, thousand yuan, ten thousand yuan, money...	Criminal outcome
Addition and Reduction of Sentences	Meritorious service, surrender, lesser amount, lesser, lessened...	Commutation
Addition and Reduction of Sentences	A huge amount, a particularly huge amount, heavy, aggravated...	Increase the sentence
Judgment result	Imprisonment, probation, fine...	Sentencing results

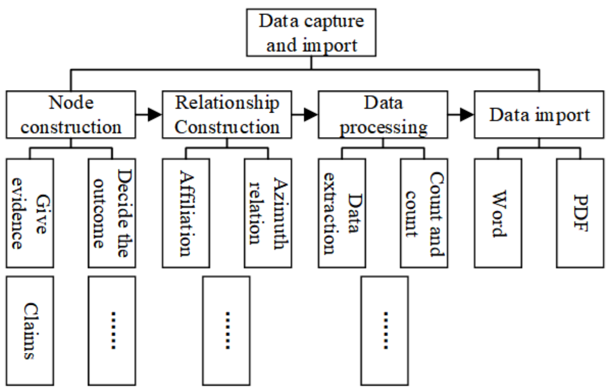

Figure 1: Data settlement and import

Since there are many invalid texts in the document, and the key vocabulary of each feature cannot be obtained at once, it is necessary to continuously enrich the vocabulary in the iterative process, so as to obtain a richer feature vocabulary and continuously improve the scalability of the program. In this paper, the method of association analysis is used to improve the accuracy of sample extraction. Establishing association rules to express the relationship between data is the goal of association analysis, which aims to validate the correlation—that is, the legal causal relationship—between objects or occurrences. Words with a strong correlation are related words when conductiof the B word appearing. Currently, the terms "A" and "B" are considered linked. We can improve the richness of the case information and identify the case's features from a vast number of documents by analyzing related words. Figure 2 illustrates the pattern matching and association analysis feature extraction procedure for case information.

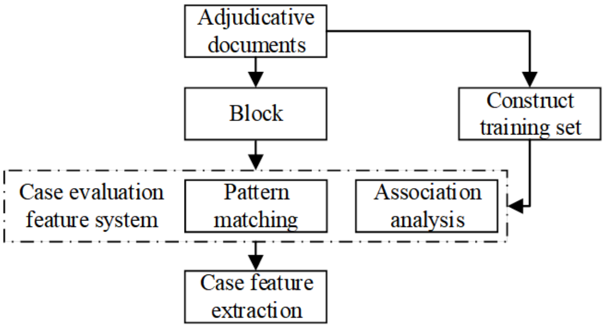

Figure 2: Case information feature extraction process for pattern matching and association analysis

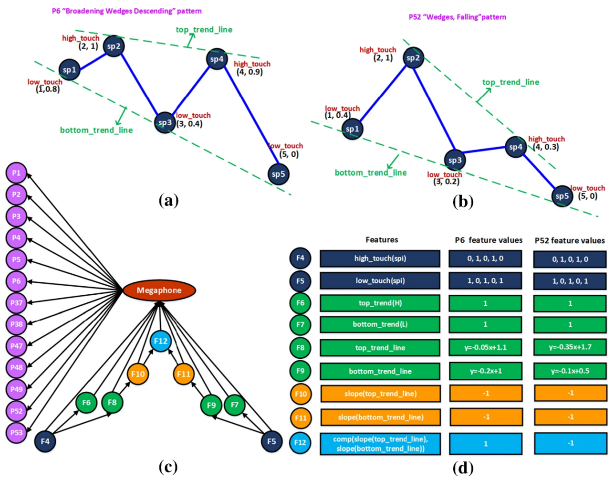

Figure 3: Pattern recognition, feature mapping, and classification network for standardized sentencing prediction

Figure 3 illustrates a visual analogy between financial pattern recognition and the proposed judicial sentencing standardization framework. Subfigures (a) and (b) demonstrate two distinct case development patterns, analogous to “Broadening Wedges Descending” and “Falling Wedges,” respectively. These patterns represent the evolving trajectories of criminal behaviors, where key nodes (e.g., high_touch, low_touch) symbolize aggravating and mitigating sentencing factors, such as intent severity or amount defrauded. SubFigure (c) shows a classification network where multiple identified patterns (P1–P53) are clustered under a unified sentencing category ("Megaphone"), akin to grouping similar legal cases under a standard sentencing guideline. This mirrors the process in our study where extracted features—such as criminal intent, means, and outcomes—are stored in a structurebase. SubFigure (d) presents the binary and numerical encoding of feature values for selected patterns, demonstrating how complex legal semantics are transformed into machine-readable inputs for LightGBM modeling. The feature dimensions (e.g., top_trend_line, slope) correspond to the severity, frequency, and progression of criminal acts, thus enabling quantitative prediction of sentencing outcomes based on structured judgment elements.

Figure 4 provides a conceptual mapping between technical pattern recognition in time-series data and standardized sentencing modeling based on judicial big data. Subfigures (a) and (b) depict two distinct trajectory patterns, labeled P7 “Bump-and-Run Reversal Bottom” and P8 “Bump-and-Run Reversal Top,” respectively. These trajectories can be interpreted as metaphors for the evolution of different categories of criminal cases over time, where key spatial points (sp1–sp7) signify critical legal elements such as criminal intent, method, and consequences. Vertical distances from touchpoints to trend lines are used to indicate the severity of aggravating or mitigating factors, analogous to features such as “particularly huge amount,” “voluntary surrender,” or “first offense” in legal sentencing. In subfigure (c), multiple criminal pattern types (e.g., P7 and P8) are grouped under a broader category “Bump-and-Run,” illustrating the classification process of diverse but structurally similar legal cases into unified smon elements across fraud cases are grouped to ensure consistent sentencing reference. subfigure (d) shows the numerical encoding of key features—including slopes between legal event nodes, durations between stages, presence of top/bottom trend lines, and vertical deviations—used to transform legal case semantics into machine-readable inputs for LightGBM prediction modeling. These encoded features correspond to the feature extraction tables and database schema described in Sections II and III of our study, ultimately supporting the construction of a quantitative and interpretable sentencing standardization model.

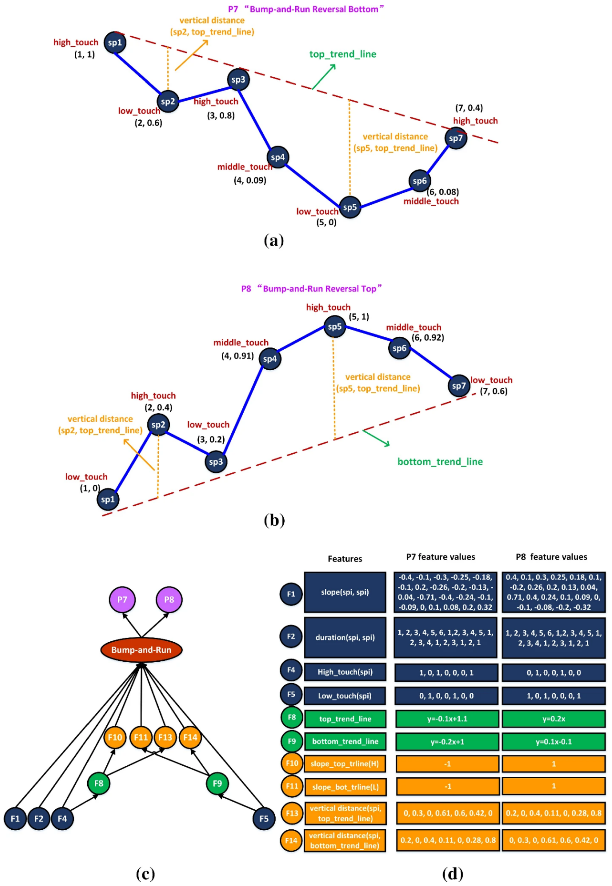

Figure 4: Structural comparison and feature encoding for "Bump-and-Run Reversal" legal case trajectories

III. The construction of the database of sentencing norms

After the features are extracted, the data needs to be stored, so it is very important to design an efficient database. Database design should be flexible and highly scalable. The database includes a case name table, a criminal basic information table, a crime purpose table, a crime means table, a crime result table, a commutation table, a sentence increase table, and a sentencing result table. The E-R diagram of the database is shown in Figure 5.

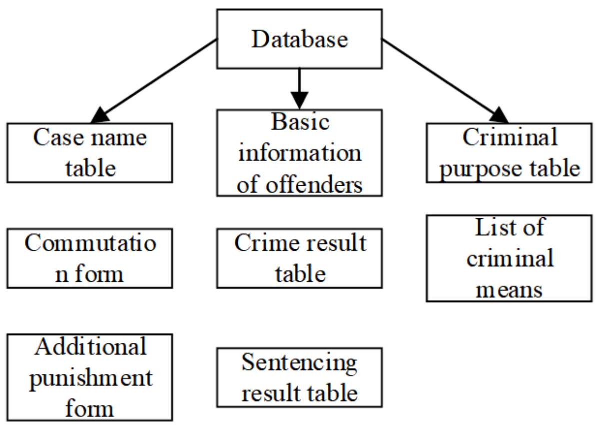

Figure 5: Database E-R diagram

The following is a description of the entity relationship in the database. The entity table of basic information of judgment documents is shown in Table 2. The table name is judgement, which mainly stores the basic information of the case, including the case number, trial procedure, court name, etc. These information will be displayed as the brief information of the case.

Table 2: Basic information entity table of judgment documents

Name	Field name	Type	Length
ID	id	int	
Case number	case_no	VARCHAR	50
Trial procedure	hear_procedure	tinyint	
Review level	trial_level	tinyint	
Instrument	uri	varchar	100
Court name	court_name	varchar	100
Mark	tag	int	
Case type	case_type	VARCHAR	50
Cause of action	cause_of_action	VARCHAR	100

The basic information table of criminals is shown in Table 3. The name of the table is criminal, which mainly stores relevant information of criminals.

Table 3: Basic information of offenders

Name	Field name	Type	Length	Remark
ID	ID	int		Table primary key, auto increment
Gender	Gender	VARCHAR	20	Relative address in the server
Name	Name	VARCHAR	20	
Identity number	ID	VARCHAR	100	
Province	Province	VARCHAR	50	
City	City	VARCHAR	50	
County/District	Region	VARCHAR	50	
Age	Age	VARCHAR	20	

The crime purpose table is shown in Table 4. The table name is intention, which mainly stores the relevant information of the crime purpose.

Table 4: Crime purpose table

Name	Field name	Type	Length
ID	id	int	
Purpose	intention	VARCHAR	100

The crime means table is shown in Table 5. The table name is means, which mainly stores the relevant information of the crime means.

Table 5: List of means of crime

Name	Field name	Type	Length
ID	id	int	
Means	means	VARCHAR	100

The crime results table is shown in Table 6. The table name is results, which mainly stores the relevant information of the crime results.

Table 6: Crime result table

Name	Field name	Type	Length
ID	id	int	
Result	results	VARCHAR	200

The crime result table is shown in Table 7. The table name is reduce, which mainly stores the relevant information of commutation.

Table 7: Commutation table

Name	Field name	Type	Length
ID	id	int	
Commutation	reduce	VARCHAR	300

The crime result table is shown in Table 8. The table name is additional, which mainly stores the relevant information of the penalty.

Table 8: Aggravation table

Name	Field name	Type	Length
ID	id	int	
Increase the sentence	additional	VARCHAR	300

Table 9: Sentencing result table

Name	Field name	Type	Length
ID	id	int	
Sentencing result table	sentencing	VARCHAR	300

The sentencing result table is shown in Table 9. The table name is sentencing, which mainly stores the relevant information of the sentencing result.

Figure 6 illustrates a structured abstraction of legal case progression using two symmetrical temporal-spatial patterns: P9 “Cup with Handle” and P10 “Cup with Handle, Inverted.” Subfigures (a) and (b) model the dynamic development of judicial cases where sequential stages (sp1 to sp7) represent distinct legal phases such as premeditation, execution, and sentencing outcome. The “handle” region—indicated by the spatial relationship and amplitude between nodes (e.g., sp3 to sp5)—can be viewed as the turning point in a case, such as the discovery of mitigating evidence or a defendant’s change in plea. Duration and amplitude calculations (e.g., duration(sp5, sp7), amplitude(sp5, sp6)) reflect the temporal spacing and intensity of case features, comparable to intervals between court appearances or degrees of criminal damage. In subfigure (c), patterns P9 and P10 are unified under a broader classification "Cup with Handle," demonstrating the generalization process by whicrent case configurations can be grouped into standard sentencing types. This mirrors the approach in our study, where fraud-related legal documents are parsed and classified based on structural and semantic similarities into consistent sentencing schemas. subfigure (d) reveals how complex legre encoded into machine-readable formats: slope, duration, and amplitude (F1–F3) represent behavioral patterns of the defendant, while top and bottom trend lines (F8–F9) serve as contextual baselines for legal assessment. Feature vectors such as slope(top_trend_line) and slope(bottom_trend_line) (F10–F11) indicate whether sentencing aggravation or leniency trends exist.

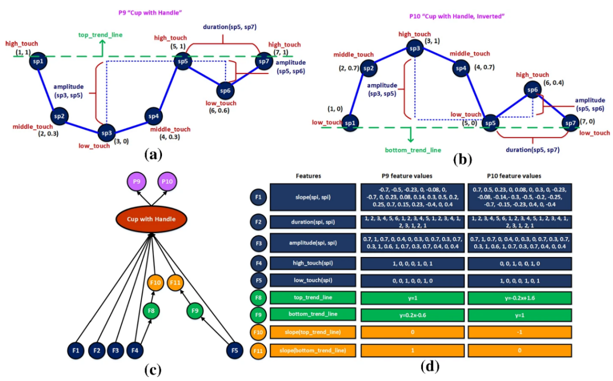

Figure 6: Semantic transformation and feature representation for “Cup with Handle” patterns in sentencing trajectory modeling

Figure 7 provides a structural abstraction of complex sentencing trajectories using the P11 “Diamond Bottoms” pattern. In subfigure (a), the sequential arrangement of critical event nodes (sp1–sp7) forms a symmetric, diamond-like structure, symbolizing legal cases with alternating aggravating and mitigating factors. Each “high_touch” and “low_touch” point corresponds to distinct legal conditions—such as crime severity spikes (e.g., organized crime involvement) or dips due to mitigating factors (e.g., voluntary confession). The geometric alignment along top and bottom trend lines mirrors the interpretive constraints within judicial sentencing: upper bounds (maximum penalty thresholds) and lower bounds (minimum statutory sentencing). In subfigure (b), the two diamond-based patterns P11 and P12 are abstracted under a unified semantic category “Diamond,” reflecting their shared legal structure despite different surface-level features. This process parallels the grouping of structurally similar fraud ca standardize sentencing. subfigure (c) displays the feature encoding results of the P11 pattern: binary values for high/low touches (F4, F5), numerical representations for trend frequencies (F6, F7), parametric expressions of upper/lower boundaries (F8, F9), and slope calculations (F10, F11) that signal increases or decreases in legal severity. These structured feature values serve as quantitative inputs to the LightGBM-based sentencing model described in Section IV, enabling precise prediction and comparison of sentencing outcomes.

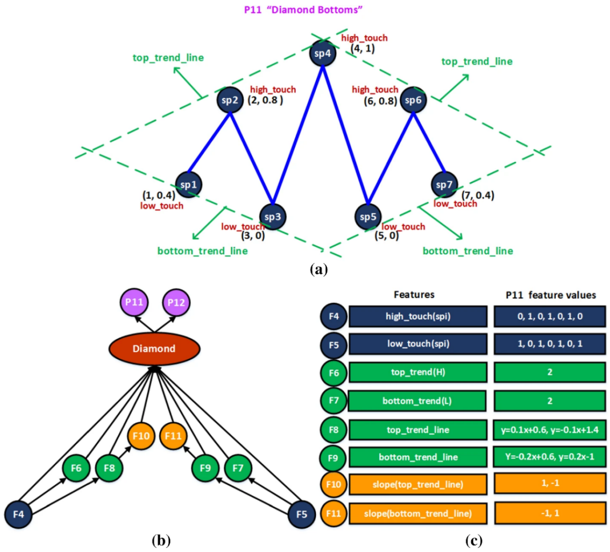

Figure 7: Feature mapping and structural symmetry in "Diamond Bottoms" sentencing patterns

IV. Sentencing standardization model design

After completing the identification of the elements of the judgment document through pattern matching and association analysis model, with all elements as the characteristics and the sentencing result as the prediction target, the LightGBM algorithm is used to establish a standardized sentencing model based on judicial big data. The label settings of sentencing results are shown in Table 10.

Table 10: Sentencing result labels

Sentencing results	Label
Probation	0
1 year imprisonment or less	1
1 year to 3 years imprisonment	2
More than three years in prison but less than five years	3
5 years to 7 years imprisonment	4
imprisonment for a minimum of seven years and a maximum of ten years	5
10 years or more in prison	6
life imprisonment	7

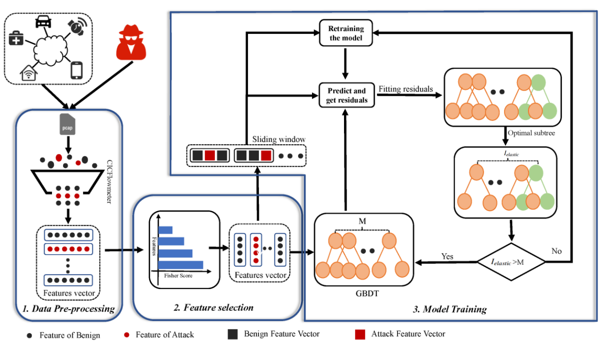

Figure 8: Workflow of judicial sentencing prediction based on feature filtering and GBDT residual feedback

Figure 8 outlines the complete workflow of the sentencing standardization model based on judicial big data, using gradient boosting decision trees (GBDT) with an iterative residual learning mechanism. In phase 1, "Data Pre-processing," legal case documents—specifically, sentencing texts—are parsed using a judicial feature extractor analogous to CICFlowmeter in cybersecurity, where judgment paragraphs are tokenized and structured into feature vectors. Each vector represents a sequence of legal attributes such as criminal intent, method, result, and judicial outcome, with "benign" and "attack" labels corresputlier sentencing patterns (e.g., disproportionate sentencing). Phase 2, "Feature Selection," applies statistical scoring such as Fisher

Score to select the most legally discriminative features (e.g., extreme amounts defrauded, aggravating circumstances). These are used to reduce dimensionality and enhance the quality of sentencing predictors. Phase 3, “Model Training,” initiates the LightGBM (GBDT) process to train on these features. A sliding window technique enables sequential case analysis—each legal case vector is compared against previously seen decisions to identify inconsistencies. Residuals are calculated (i.e., the difference between predicted and actual sentencing outcomes), and the model is retrained on these residuals to refine prediction accuracy.

The model evaluates the optimality of each decision subtree based on an elastic loss function (I_elastic), reflecting legal interpretability and sentencing consistency. If the new subtree improves upon the base model (I_elastic $\geq$ M), the model structure is updated. Otherwise, the original configuration is retained, ensuring stability in judgment prediction.

V. Experiments

Randomly select 1,000 judgment documents of online fraud cases from the Judgment Documents Network, and extract the information features of the cases to convert the judgments into data, and then perform calculations based on the standardized sentencing modse (Table 11), the comparison between the average value of the sentence calculated by the model and the average value of the actual sentence of the case is poor (Table 12), which proves that the normalized sentencing model has good measurement ability.

Table 11: The difference between the estimated sentence of the model and the actual sentence (select 20 cases)

Cases (1000 cases in total, 20 cases selected)	Model Estimates Sentences	Actual sentence	Relatively poor
XX People’s Court Criminal Judgment	1	1	0
XX People’s Court Criminal Judgment	3	3	0
XX People’s Court Criminal Judgment	4	5	1
XX People’s Court Criminal Judgment	10	10	0
XX People’s Court Criminal Judgment	5	5	0
XX People’s Court Criminal Judgment	3	4	1
XX People’s Court Criminal Judgment	3	3	0
XX People’s Court Criminal Judgment	6	7	1
XX People’s Court Criminal Judgment	6	6	0
XX People’s Court Criminal Judgment	3	3	0
XX People’s Court Criminal Judgment	5	4	1
XX People’s Court Criminal Judgment	6	6	0
XX People’s Court Criminal Judgment	8	9	1
XX People’s Court Criminal Judgment	3	3	0
XX People’s Court Criminal Judgment	10	11	1
XX People’s Court Criminal Judgment	9	9	0
XX People’s Court Criminal Judgment	10	11	1
XX People’s Court Criminal Judgment	10	10	0
XX People’s Court Criminal Judgment	15	15	0
XX People’s Court Criminal Judgment	15	13	2

Table 12: The discrepancy between the model’s estimated average sentence value and the actual sentence’s average

Case	The average value of the sentence is determined by the model.	Average actual sentence	Relatively poor
Internet Fraud Cases	65.76	64	1.76

The sentencing standardization model framework based on judicial big data adopted by this research can be used to predict and judge sentencing standardization. In the future, the judicial field can design different feature extraction models according to the characteristics of different case types, conduct automatic classification and data conversion of court judgments, use statistical linear regression method to analyze court judgment data in real time, and establish sentencing databases for various cases to improve the real-time service of sentencing databases.

Figure 9 presents a comparative analysis of sentencing prediction accuracy across seven distinct models (Model I to Model VII), evaluating three machine learning approaches—Grid-XGBoost, Gradient Tree Boosting (GTB), and Random Forest (RF)—against actual observed sentencing data. Each subplot displays a boxplot distribution of predicted sentence severity (in thousands of units, Kx), highlighting the central tendency, variance, and outliers associated with each model’s predictions. The “Observed” box represents the true sentencing distribution derived from judicial documents, serving as the baseline for model comparison. From a judicial perspective, these boxplots assess the consistency and faof each machine learning model in replicating sentencing decisions based on legal features such as criminal intent, harm severity, and mitigating/aggravating factors. For instance, models showing close alignment with the “Observed” distribution (e.g., Grid-XGBoost in Model III and Model VI) demonstrate superior capacity for understanding judicial logic and normatively grounded sentencing ranges. Conversely, wider spreads or skewed medians indicate prediction instability or misalignment with legal precedents. This form of evaluation is crucial in judicial applications, as predictive accuracy is not merely a matter of numerical closeness but also of interpretive fidelity to the principles of proportionality and justice. Outlier detection (e.g., sentences exceeding 750 Kx) helps identify cases where model outputs deviate significantly from established norms, possibly flagging inconsistencies due to model bias or underfitting in complex sentencing scenarios.

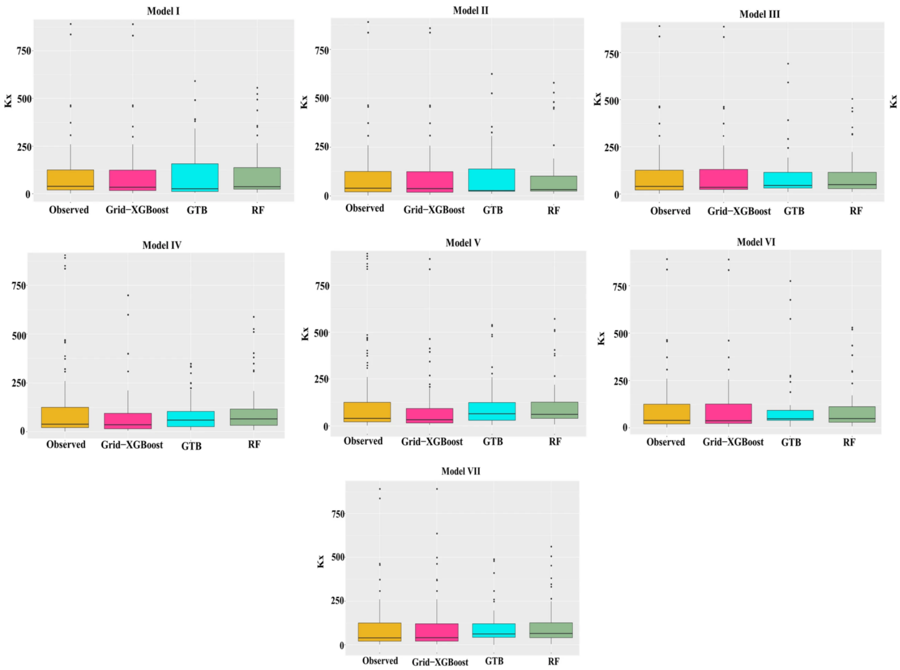

Figure 9: Boxplot comparison of observed versus predicted sentencing results across seven judicial prediction models using Grid-XGBoost, GTB, and RF

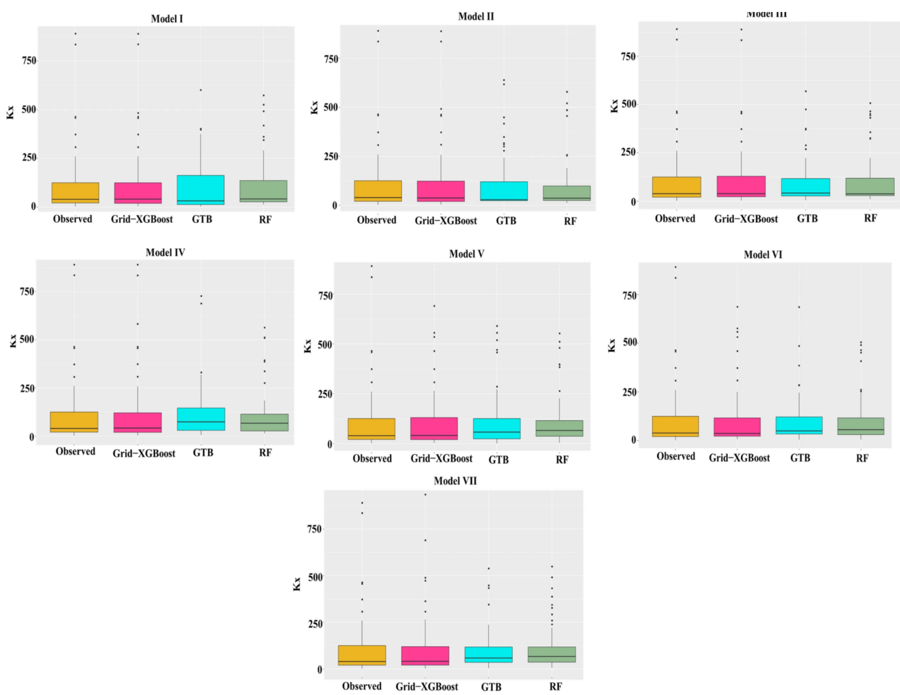

Figure 10: Comparative boxplot analysis of observed sentencing versus model predictions under seven legal modeling scenarios

Figure 10 provides a comprehensive boxplot comparison between actual observed sentencing outcomes and the predictions generated by three machine learning models—Grid-optimized XGBoost, Gradient Tree Boosting (GTB), and Random Forest

(RF)—across seven distinct model configurations (Model I to Model VII). Each subplot captures the statistical distribution of sentencing values in terms of thousand-unit scales (Kx), enabling visual evaluation of central tendency, dispersion, and outlier behavior. These models were trained on structured legal features extracted from fraud-related judicial documents, including crime severity, economic loss, intent level, and mitigating or aggravating model's capacity to replicate real-world sentencing distribution—crucial for supporting judicial fairness, consistency, and data-driven adjudication. In general, the predicted distributions from Grid-XGBoost most closely align with the observed sentencing medians and interquartile ranges, particularly in Model IV and Model VII. This suggests that Grid-XGBoost not only minimizes prediction variance but also reflects better interpretability and conformity to normative legal standards. RF and GTB, while stable, occasionally exhibit broader spread or downward-biased medians, indicating underestimation risks in complex sentencing cases.

Outliers in both observed and predicted results reflect rare but severe legal penalties (e.g., life imprisonment or high-value fraud), and their alignment across models shows that the feature design effectively captures extreme sentencing scenarios.

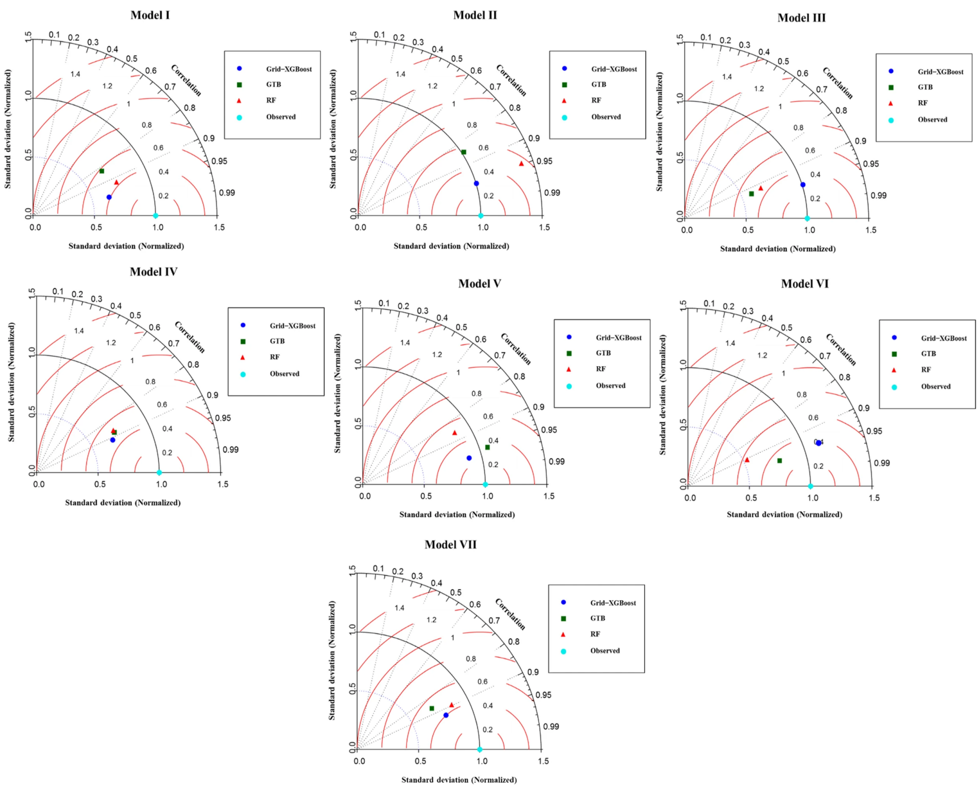

Figure 11: Taylor diagram comparison of observed and predicted sentencing outcomes across seven model configurations

Figure 11 presents Taylor diagrams for seven sentencing prediction models, assessing the performance of Grid-optimized XGBoost, Gradient Tree Boosting (GTB), and Random Forest (RF) against actual observed sentencing results. Each subfigure quantifies model similarity using three core indicators: correlation coefficient, normalized standard deviation, and centered root-mean-square difference. In the context of judicial analytics, these metrics collectively represent how well each model replicates the complexity, spread, and structure of real-world sentencing outcomes. The turquoise point in each diagram corresponds to the observed data, while the proximity of model markers (blue: Grid-XGBoost, green: GTB, red: RF) to this reference point indicates predictive fidelity. Across all models, Grid-XGBoost consistently exhibits the highest correlation (0.95–0.99) with the observed data, and its standard deviation is closest to 1.0—indicating it preserves the legal sentencing variance without overfitting or oversimplification. GTB and RF occasionally show greater standard deviations or lower correlations, suggesting slightly less stability in capturing nuanced legal decision patterns.

Figure 12 presents a set of Taylor diagrams comparing the predictive performance of Grid-XGBoost, Gradient Tree Boosting (GTB), and Random Forest (RF) models across seven sentencing prediction scenarios. Each subplot represents one model configuration, where the positioning of points along the radial and angular dimensions reflects the model's normalized standard deviation and correlation with observed sentencing outcomes.The turquoise circles denote the baseline distribution from real-world sentencing data. The blue (Grid-XGBoost), green (GTB), and red (RF) markers indicate edel's performance relative to this benchmark. A model point that is closest to the observed point with a correlation near 1.0 and a standard deviation near

1.0 signifies a highly interpretable, norm-aligned prediction structure—a critical requirement in legal AI applications where fairness and consistency are paramount.

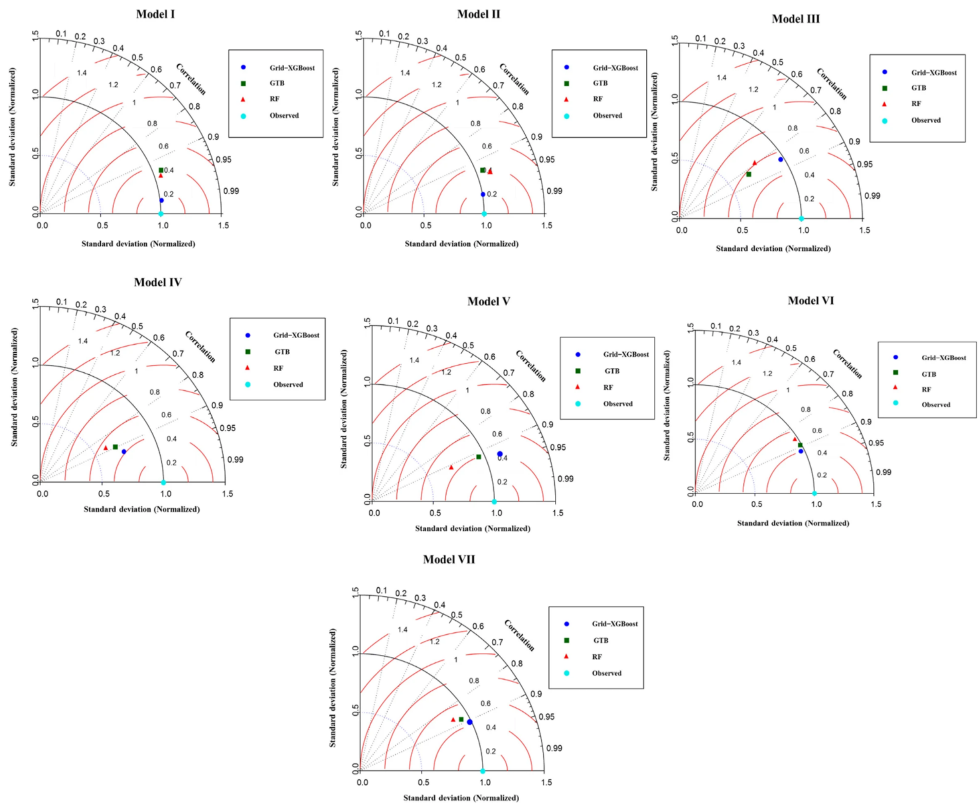

Figure 12: Taylor diagram comparison of normalized standard deviation and correlation for sentencing prediction models

Across most scenarios—especially in Models I, IV, and VI—Grid-XGBoost shows superior alignment, residing closest to the ideal observed point. It maintains low root-mean-square errors, high correlation (>0.95), and standard deviation values tightly clustered around 1.0, suggesting both statistical accuracy and legal structural fidelity. By contrast, RF and GTB show slightly more dispersed placements, with greater deviation from legal sentencing variability, potentially reflecting model underfitting or lack of discrimination in edge-case legal conditions (e.g., mixed aggravating/mitigating circumstances).

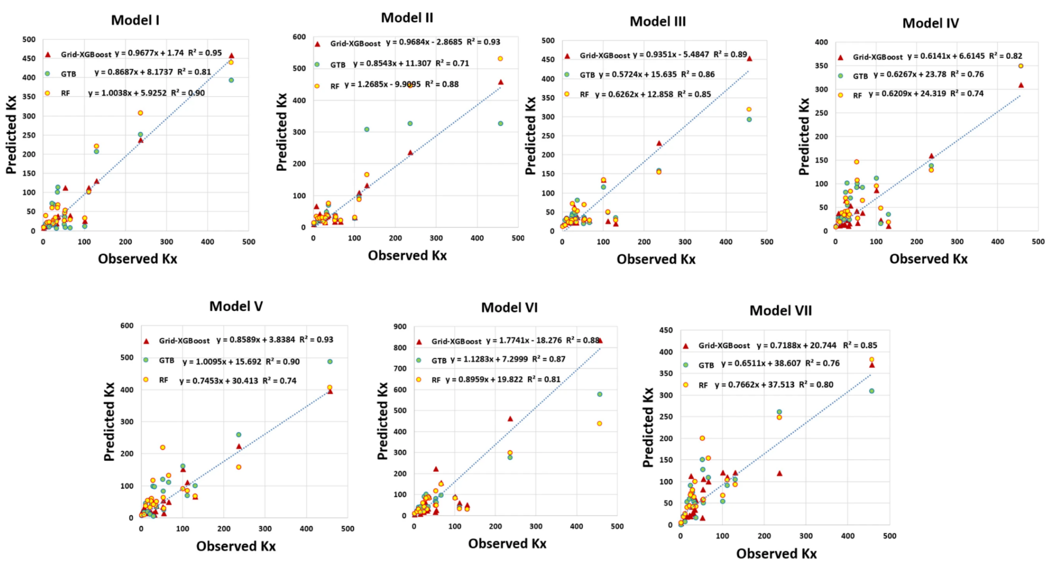

Figure 13: Regression fitting analysis between observed and predicted sentencing results across seven model settings

Figure 13 presents scatter plots of observed sentencing values (Kx) versus model-predicted sentencing outcomes using Grid-optimized XGBoost (red triangles), Gradient Tree Boosting (green circles), and Random Forest (orange dots) across seven distinct model settings (Model I–VII). Each plot includes the best-fit regression line, linear equation, and coefficient of determination (R^2), which collectivcomes. From a legal analytics perspective, a high R^2 value combined with a slope close to 1.0 indicates that the model not only fits the observed data well but also preserves the proportionality essential to sentencing fairness. In nearly all configurations—particularly in Models I, II, and V—Grid-XGBoost outperforms other models with R^2 values consistently above 0.90, and regression slopes near unity. This demonstrates that Grid-XGBoost maintains a high level of predictive fidelity, enabling it to reflect judicial patterns such as proportional sentencing for different levels of crime severity or damage.

In contrast, models like RF and GTB occasionally exhibit lower slopes or increased intercept offsets (e.g., Model IV and VII), suggesting they may underpredict or overpredict sentencing values in cases with nuanced legal features. For instance, RF in Model V shows an underfitting trend, predicting values significantly lower for high-observed cases (slope = 0.7453), which could misrepresent legally severe cases like aggravated fraud.

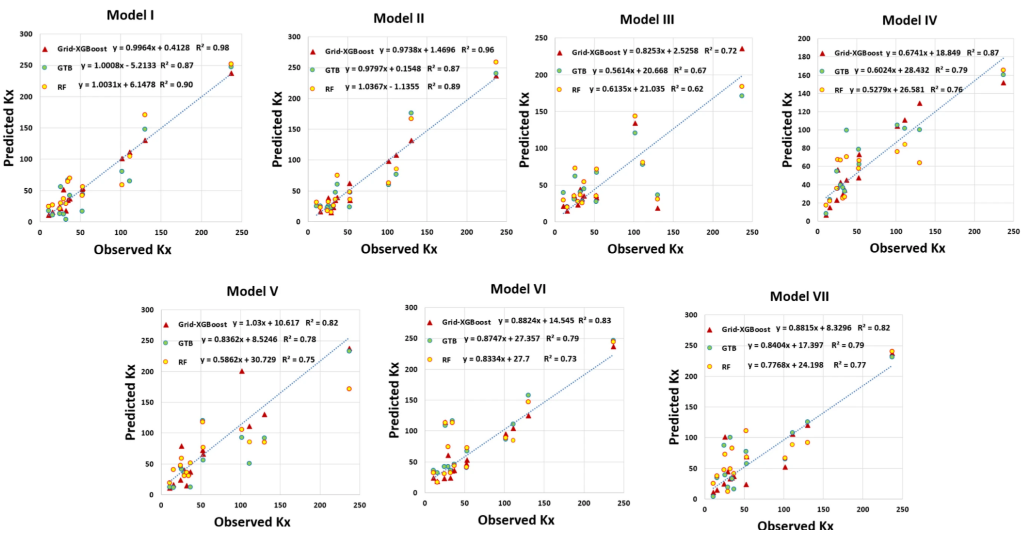

Figure 14: Regression performance analysis of sentencing prediction models under reduced case complexity (Kx < 250)

Figure 14 illustrates the regression performance of three machine learning models—Grid-optimized XGBoost, Gradient Tree Boosting (GTB), and Random Forest (RF)—across seven sentencing prediction configurations, with the observed sentencing values (Kx) constrained to a normalized legal range below 250. Each subplot shows a scatter plot of predicted versus observed Kx values, with regressidetermination (R^2). The Grid-XGBoost model maintains consistently strong performance across all seven scenarios, achieving the highest R^2 values—up to 0.98 in Model I and above 0.82 in five of the seven cases—indicating high precision in capturing sentencing variation. This performance aligns with the goal of achieving sentencing standardization, as the model closely approximates judicial reasoning while minimizing prediction deviation. Its regression slopes remain near 1.0 (e.g., 0.9964 in Model I), reflecting strong proportional mapping between legal case input features and sentencing outcomes.

GTB and RF, though effective in certain scenarios, demonstrate reduced consistency in models III and IV, where underfitting is evident (e.g., RF $R^2 = 0.62$ in Model III). These discrepancies highlight the challenge of modeling intermediate-complexity cases, where legal circumstances (e.g., partial restitution, minor accomplices) introduce interpretative nuances not easily captured by tree-based heuristics without fine-tuned feature interactions.

VI. Conclusion

This study presents a structured and scalable sentencing standardization model grounded in judicial big data and machine learning. By extracting interpretable features from real-world online fraud judgment documents—such as intent, means, severity, and judicial outcomes—and encoding them into a feature-rich database schema, we developed a LightGBM-based sentencing prediction engine that demonstrated robust align with actual judicial behavior. The experimental results—validated through statistical analyses including regression fit, Taylor diagrams, and boxplots—confirm the model's capability to accurately simulate judicial discretion and sentencing proportionality.

Beyond technical accuracy, the model also upholds legal interpretability, a crucial aspect often overlooked in algorithmic law. Through visual abstractions like "Cup with Handle" and "Diamond Bottoms," we demonstrated how legal trajectories can be structurally encoded and classified into standard sentencing archetypes. Such representations not only enhance model explainability but also align machine predictions with doctrinal expectations of fairness, consistency, and proportionality.

Data availability

The figures and tables used to support the findings of this study are included in the article.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Funding Statement

This work was not supported by any funds.

References

[1] Hasegawa, F. , Taira, A. , Noh, G. , Bing, H. , & Kim, I. . (2018). High-speed train communications standardization in 3gpp 5g nr. *IEEE Communications Standards Magazine, 2*(1), 44-52.

[2] Chak-Yu, S. , Shing-Him, C. G. , Chan, K. Y. , Pui-Wai, L. A. , & Yat-Yin, L. . (2018). A call for standardization in left atrial appendage occlusion. *Journal of the American College of Cardiology, 72*(4), 472-473.

[3] Xu, Y., Tao, Y., Zhang, C., Xie, M., Li, W., & Tai, J. (2022). Review of digital economy research in China: a framework analysis based on bibliometrics. *Computational Intelligence and Neuroscience, 2022*(1), 2427034.

[4] Golasi, I. , Salata, F. , Vollaro, E. D. L. , & Coppi, M. . (2018). Complying with the demand of standardization in outdoor thermal comfort: a first approach to the global outdoor comfort index (goci). *Building & Environment, 130*(feb.), 104-119.

[5] Wang Lusheng. (2018). Research on discourse conflict in the practice of judicial big data and artificial intelligence. *Legal Forum, 033*(005), 137-144.

[6] Zhu, F. B. , & Song, Z. . (2022). Systematic regulation of personal information rights in the era of big data. *SAGE Open, 12*(1), 222-229.

[7] Milaninia, N. . (2020). Biases in machine learning models and big data analytics: the international criminal and humanitarian law implications. *International Review of the Red Cross, 102*(913), 199-234.

[8] Sun, S. , & Sun, Y. . (2020). Empirical study of the dilemma regarding preliminary proof of causality and methods used by victims in china's environmental health litigation. *China's Population, Resources and Environment: English Edition, 18*(3) , 7.

[9] Engin, Z. , & Treleaven, P. . (2019). Algorithmic government: automating public services and supporting civil servants in using data science technologies. *The Computer Journal, 62*(3), 448-460.

[10] Nissi, E. , Cusatelli, C. , & Maggino, F. . (2019). The efficiency of the italian judicial system: a two stage data envelopment analysis approach. *Social Indicators Research, 146*(1-2), 395-407.

[11] Ni, B. . (2020). Research on risk prevention system of judicial administration based on big data analysis. *Journal of Physics: Conference Series, 1648*(4), 042074 (8pp).

[12] Xu, Z. (2022). Human judges in the era of artificial intelligence: challenges and opportunities. *Applied Artificial Intelligence*, *36*(1), 2013652.

[13] Ryo, E., & Humphrey, R. (2022). The importance of race, gender, and religion in naturalization adjudication in the United States. *Proceedings of the National Academy of Sciences*, *119*(9), e2114430119.

[14] Kennedy, P. (2022). Diffusion of Soft Immigration Law: Evidence from Asylum Adjudication in the Wake of Matter of AB. *Montana Law Review*, *83*(1), 2.

[15] Dallara, C., Lacchei, A., & Verzelloni, L. (2022). Up-scaling social innovation in asylum adjudication: the case of the Migrantes project in Sicily. *Innovation: The European Journal of Social Science Research*, 1-20.