Artificial Intelligence and Law
https://doi.org/10.1007/s10506-023-09362-x

ORIGINAL RESEARCH

LK-IB: a hybrid framework with legal knowledge injection for compulsory measure prediction

Xiang Zhou[1] · Qi Liu[1] · Yiquan Wu[2] · Qiangchao Chen[1] · Kun Kuang[2]

Accepted: 16 May 2023

Abstract

The interpretability of AI is just as important as its performance. In the LegalAI field, there have been efforts to enhance the interpretability of models, but a tradeoff between interpretability and prediction accuracy remains inevitable. In this paper, we introduce a novel framework called LK-IB for compulsory measure prediction (CMP), one of the critical tasks in LegalAI. LK-IB leverages Legal Knowledge and combines an Interpretable model and a Black-box model to balance interpretability and prediction performance. Specifically, LK-IB involves three steps: (1) inputting cases into the first module, where first-order logic (FOL) rules are used to make predictions and output them directly if possible; (2) sending cases to the second module if FOL rules are not applicable, where a case distributor categorizes them as either "simple" or "complex"; and (3) sending simple cases to an interpretable model with strong interpretability and complex cases to a black-box model with outstanding performance. Experimental results demonstrate that the LK-IB framework provides more interpretable and accurate predictions than other state-of-the-art models. Given that the majority of cases in LegalAI are simple, the idea of model combination has significant potential for practical applications.

Keywords Legal knowledge · Model combination · Compulsory measure prediction · Interpretability

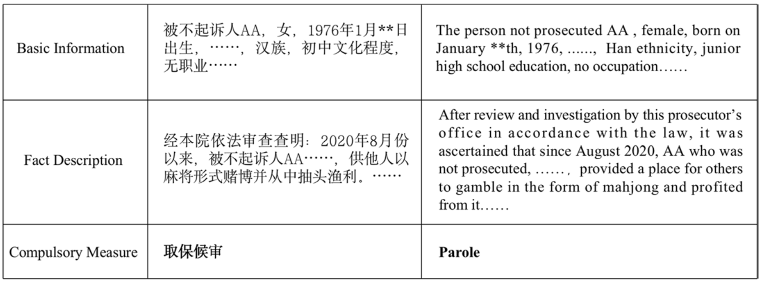
1 Introduction

In recent years, Legal Artificial Intelligence (LegalAI) has emerged as a crucial tool for reducing the heavy and repetitive workload in the legal field. By leveraging artificial intelligence technology, LegalAI can assist with various legal tasks, including judgment prediction, legal question answering, and more Zhong et al. (2020). For example, previous studies have explored the use of LegalAI for judgment

Xiang Zhou and Qi Liu contributed equally and are co-first authors.

Extended author information available on the last page of the article

Published online: 30 May 2023



Basic Information	被不起诉人AA，女，1976年1月**日出生，……，汉族，初中文化程度，无职业……	The person not prosecuted AA , female, born on January **th, 1976, ..., Han ethnicity, junior high school education, no occupation……
Fact Description	经本院依法审查查明：2020年8月份以来，被不起诉人AA……，供他人以麻将形式赌博并从中抽头渔利。……	After review and investigation by this prosecutor's office in accordance with the law, it was ascertained that since August 2020, AA who was not prosecuted, ……， provided a place for others to gamble in the form of mahjong and profited from it……
Compulsory Measure	**取保候审**	**Parole**

Fig. 1 An example of CMP. Given the basic information and the fact description, the prosecutor should determine the compulsory measure. The three components can be objectively extracted from a legal document with rules

prediction Liu and Chen (2018); Luo et al. (2017); Long et al. (2019) and legal question answering Kim and Goebel (2017); Do et al. (2017); Fawei et al. (2019). These efforts have demonstrated the potential of LegalAI to enhance the efficiency and accuracy of legal processes.

Compulsory measure prediction (CMP) is a critical task in LegalAI as it aims to assist prosecutors (or judges, depending on the country) in determining whether a suspect should be detained or paroled. Detention is one of the harshest compulsory measures that can be imposed on a suspect, and it can result in a significant infringement of their personal freedom without trial. To safeguard the fundamental human rights of suspects, several research efforts have been conducted, such as the COMPAS Brennan et al. (2009) and FPRAI Cohen and Lowenkamp (2018) systems in the US. However, the task of CMP encounters different case information structures in different countries due to variations in legal systems Zhou et al. (2022). Therefore, methods designed for one country cannot be readily applied to another, highlighting the need for tailored approaches for each country.

China, being a heavily populated country, has a significant number of legal cases, making it imperative to assist prosecutors in determining the appropriate compulsory measure. However, there are limited studies relevant to CMP based on Chinese legal datasets. Therefore, this paper aims to fill this blank.

As depicted in Fig. 1, determining the appropriate compulsory measure requires prosecutors to take into account both the basic information of the suspect and the facts of the case. The suspect's basic information usually includes criminal history and behavior after the crime. For instance, if there is evidence suggesting that the suspect intends to commit additional crimes or flee, the prosecutor may opt for detention. On the other hand, non-custodial measures such as parole may be more appropriate for monitoring suspects when the crime is less severe. In contrast, in charge prediction tasks, which are another important aspect of LegalAI, basic information is typically disregarded.

The task of CMP can be abstracted as a text classification task, which is a classic problem in Natural Language Processing (NLP). However, it is crucial that the output of CMP is interpretable, as imposing compulsory measures on suspects requires sufficient justification for both the individuals involved and the general public.

Firstly, adopting compulsory measures that restrict personal freedom, such as detention, has the potential to violate the fundamental human rights of suspects. Therefore, it is essential to exercise caution when making decisions regarding such measures. Secondly, when predicting the type of compulsory measure, it is crucial to take into account the social risks posed by the suspects. This ensures that the measures imposed are appropriate and necessary for ensuring public safety.

In fact, because legal decisions have significant implications for human rights, there is an increasing need to improve the interpretability of LegalAI systems. To this end, researchers are now focusing on developing effective solutions to explain the predictions of these systems, rather than solely maximizing prediction accuracy Gan et al. (2021); Jiang et al. (2018); Ye et al. (2018). However, there exists a trade-off between explanation quality and prediction accuracy Wang and Lin (2021).

Therefore, we face two main challenges when designing a CMP model: (1) How can a highly accurate prediction model be developed that is suitable for the Chinese dataset, and (2) How can we improve the model's interpretability without sacrificing too much accuracy?

In judicial practice, cases are often categorized as simple or complex by prosecutors before making legal decisions. Prosecutors evaluate the simplicity of a case based on several criteria, such as the clarity of facts and the possibility of direct judgment in accordance with the law without extensive professional expertise. Simple cases can be resolved expeditiously, while complex cases require careful analysis. By distinguishing between simple and complex cases and handling them separately, efficiency can be improved, and the attention of case officers can be focused on complex cases.

Inspired by the model combinations Wang and Lin (2021); Madras et al. (2018) and the prosecutors' decision-making process, in order to address the above challenges, we design a novel hybrid framework LK-IB by incorporating Legal Knowledge and combining Interpretable and Black-box models to balance the accuracy and model interpretability in CMP. To represent the legal knowledge, we use the first-order logic (FOL); to combine the interpretable model and black-box model, we use a case distributor to categorize whether the case is "simple" or "complex“. Specifically, our framework consists of three steps: (1) if FOL rules can make a prediction, the result will be output directly. (2) if not, the case distributor will send the case to the interpretable model or the black-box model. If the case is determined as a "simple case“, it will be sent to the interpretable model; otherwise, it will be determined as a "complex case" and will be sent to the black-box model. (3) The chosen model will make the prediction.

Meanwhile, we collect legal documents of indictments published on the official website of the Supreme People's Prosecutor of China[1] and construct a real-world dataset to validate this work.

The corpus used in this work is released by Chinese public authorities and has been anonymized wherever necessary. Therefore, our dataset does not involve any personal privacy. The experimental results on this real-world dataset show

[1] https://www.12309.gov.cn.

that LK-IB achieves excellent accuracy compared to the state-of-the-art prediction models while gaining better interpretability. Besides, intelligent prediction of legal decisions remains a sensitive technology, so it is worthwhile to investigate certain ethical considerations. The various methods currently proposed in China's LegalAI field aim to reduce the workload for legal professionals and improve their work efficiency rather than replacing humans Zhou et al. (2022); Bi et al. (2022); Wu et al. (2020). We make a detailed discussion in Sect. 5.

Our main contributions are summarized as follows:

1. We investigated the problem of compulsory measure prediction (CMP) in China from the perspective of interpretability, and introduced the idea of model combination into the CMP. As far as we know, no previous work did address the problem in the Chinese dataset.
2. We proposed a novel hybrid framework with legal knowledge injection (LK-IB) by simulating the prosecutors in the decision-making process. The LK-IB framework firstly checks if the first-order logic (FOL) ruled can make the decision of a case, if not, it uses a case distributor to categorize whether the case is "simple" or "complex" and send the case to the interpretable model or the black-box model, respectively. Finally, the selected model will make the prediction.
3. We constructed a new dataset whose contents are produced by Chinese procuratorial authorities. Experiments on the dataset validate the effectiveness of our proposed method in terms of both automatic metrics and human evaluation.
4. To motivate other scholars to investigate this novel but important problem, we will release the code and data.

The paper is structured as follows. Section 2 discusses the related work. The detail of our approach is introduced in Sect. 3. And in Sect. 4 we describe the dataset we constructed, and then we describe the experiments that we have conducted for this study and report the results. Furthermore, we make an ethical discussion in Sect. 5. Finally, we conclude our work in Sect. 6.

2 Related work

2.1 Legal artificial intelligence

Legal artificial intelligence (LegalAI) has become an essential part of legal work, which aims to bring convenience to legal professionals and the general public. Many researchers have devoted considerable efforts to promote the development of LegalAI, and the tasks of LegalAI include: Legal Judgement Prediciton(LJP) Liu and Chen (2018); Luo et al. (2017); Long et al. (2019), Court View Generation Li and Zhang (2021), Legal Question Answering Kim and Goebel (2017); Do et al. (2017); Fawei et al. (2019) and so on. LJP task plays a key role in LegalAI, which aims to predict a legal case's judgment based on a given text describing the facts of the case. The LJP task usually includes three subtasks: charges prediction, legal provisions recommendation and sentencing prediction Xu et al. (2020).

2.2 Compulsory measure prediction

Compulsory measure prediction (CMP) aims to assist prosecutors (judges in some countries) to determine whether a suspect should be detained or paroled. Although it is also a text classification task, unlike the LJP task that usually only takes the fact description as the input, the compulsory measure prediction task requires the basic information and criminal history of the suspects as input as well.

Since the 1980s, there are worldwide research on quantitative assessment of the social danger of suspects Dionne (2013). In various European and American countries, the use of risk assessment instruments as an important basis for pre-trial decision-making has gradually become widespread Desmarais et al. (2021), and can be viewed as a special type of CMP task. For instance, in the United States, the decision to implement pre-trial detention on suspects is made by judges, and the COMPAS system Dieterich (2010) can assist judges in making pre-trial decisions by predicting the probability that a suspect will commit another crime. However, China's pre-trial detention system differs significantly from those of European and American countries in several aspects. In China, compulsory measures are made by the prosecutor, with judges serving as reviewers of the legality of the detention. Moreover, there is no mature quantitative assessment system in China. Thus, the CMP task is an innovation based on the unique judicial background of China, and therefore cannot be directly compared with existing methods in other countries.

The task of CMP is the preorder task of LJP task, and pretrial detention has the potential to undermine the basic human rights of the suspect, so the interpretability of the prediction becomes more important.

2.3 Interpretability in LegalAI

Nowadays, AI is contributing to accelerating the shift towards a more convenient and intelligent society. However, many AI-based systems are characterized by a black-box nature, which means that their predictions cannot be easily explained. This issue has given rise to a new field of research called explainable AI (XAI) Arrieta et al. (2020), which aims to make the results of AI systems more understandable to humans Adadi and Berrada (2018); Adler et al. (2016). Existing interpretability methods can be broadly classified into two types: intrinsic and post-hoc, depending on whether the interpretability method can be applied retroactively or proactively (Madsen et al. 2021). Intrinsic methods are inherently comprehensible to humans, and models with intrinsic interpretability are often referred to as “interpretable models” or “white-box models“. XAI has the potential to bring significant benefits to a wide range of domains where interpretability is urgently needed. For instance, risk assessment tools have been used to predict an offender's risk of future criminal behavior Miron et al. (2021); Singh and Mohapatra (2021). However, these tools often rely on algorithmic and data-driven decision-making, which has raised concerns among the public regarding their lack of explainability and reliability (Peeters and Schuilenburg 2018; Chugh 2021).

There have been efforts to make LegalAI more explainable in recent works Yang et al. (2020); Liu et al. (2021). One approach involves incorporating first-order logic rules, which are comprehensible and often derived from domain knowledge, into the model. Another approach is to use attention mechanisms that generate attention weights, providing insights into the neural network's reasoning behind its predictions. These methods are often combined with black-box models to improve interpretability. For example, Leilei Gan et al. Gan et al. (2021) represent legal knowledge as a set of first-order logic rules and integrate them into a co-attention network, making the judgment prediction model more interpretable. Qiaoben Bao et al. Bao et al. (2019) propose an attention-based neural model that uses an attention matrix calculated from relevant articles to filter out irrelevant information in fact descriptions. Nuo Xu et al. Xu et al. (2020) presents a novel attention mechanism for extracting key features to distinguish confusing law articles attentively. Currently, the concept of "knowledge-driven & data-driven" has emerged, and the above-mentioned works demonstrate that injecting legal knowledge can significantly enhance the performance of LegalAI systems.

2.4 Hybrid models

Though many methods have been proposed to enhance our understanding of black-box models, they often result in a degradation of prediction accuracy to varying degrees Branting et al. (2021); Hacker et al. (2020). In fact, intelligent systems still face the challenge of balancing interpretability and prediction performance, particularly accuracy. However, there are a few works that focus on combining multiple models to tackle this dilemma. The core idea of hybrid models is to integrate an interpretable model with a pre-trained black-box model to leverage their strengths. David Madras et al. proposed a joint decision-making framework that generalizes rejection learning by considering the influence of other models in the decision-making process Madras et al. (2018). This framework has been shown to significantly improve the accuracy and fairness of the entire system. Tong Wang et al. proposed a hybrid predictive model and designed training algorithms to identify a subset of the data space where the interpretable model can substitute the black-box model, thereby achieving explainability at an acceptable cost in terms of prediction performance Wang and Lin (2021). These works exemplify how combining interpretable and black-box models can yield promising results in mitigating the trade-off between interpretability and prediction accuracy. Further research in this area is warranted to advance the development of hybrid models and their potential applications in real-world intelligent systems.

In real-world scenarios, there are cases where the type of necessary measures can be easily judged and considered as "simple cases" with obvious features, making them amenable to case diversion procedures. Our LK-IB is also partially inspired by the concept of hybrid models.

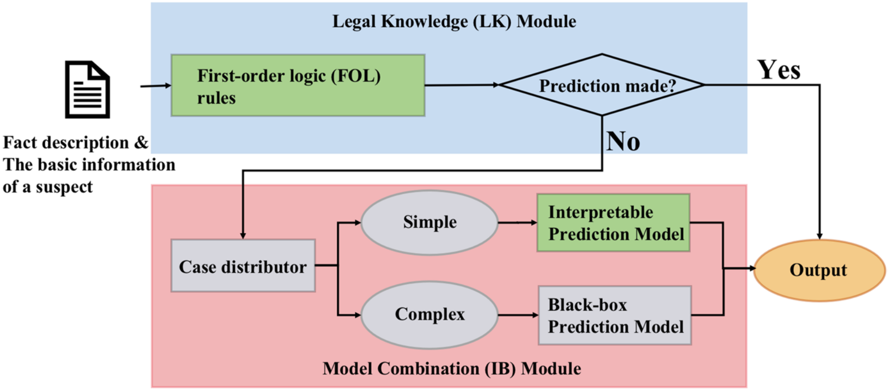

Legal Knowledge (LK) Module
First-order logic (FOL) rules
Prediction made?
Yes
No
Fact description & The basic information of a suspect
Case distributor
Simple
Complex
Interpretable Prediction Model
Black-box Prediction Model
Output
Model Combination (IB) Module

Fig. 2 Our LK-IB framework, which consists of two modules: Legal Knowledge Module and Model Combination Module

3 LK-IB framework

In this section, we introduce our novel hybrid framework, named LK-IB, which stands for a framework that injects Legal Knowledge and combines an Interpretable model with a Black box model, as illustrated in Fig. 2. The proposed framework is designed to address the challenge in the CMP task. In judicial practice, cases are often categorized as simple or complex by prosecutors before making legal decisions. Our approach is based on this observation and aims to improve interpretability without sacrificing prediction accuracy by differentiating between simple and complex cases. Interpretable models are sufficient to achieve high accuracy in predicting simple cases. Utilizing interpretable models for predicting simple cases can also enhance the interpretability of the prediction results. On the contrary, complex cases are more challenging to predict, necessitating the use of models with more complex algorithms to improve prediction accuracy.

We use the suspect's basic information and case description as input, which is a word sequence $x = \left[x_1, x_2, ..., x_n\right]$, and the compulsory measure label y ($y\epsilon\{0, 1\}$) is a non-negative integer. Given x, we aim to predict y, where $y = 0$ represents parole and $y = 1$ represents detention.

Our approach comprises two modules LK and IB, representing the legal knowledge injection and the model combination, respectively. For the first LK module, we use the First-Order Logic (FOL) rules to represent legal knowledge from the legal expert (e.g., the decision-making experience). The advantage of the LK module is that the prediction is based on the rules, so it's accurate and explainable. Since it's costly to design the FOL rules, it's impossible to cover all the possible situations. In this paper, we pick the two most frequent charges as an attempt. For the cases not suitable for the LK module, we use the second IB module to make the prediction. The IB module consists of a case distributor and two predictive models, one is an interpretable model (e.g., logistic regression) and the other is a black-box model (e.g., deep neural network).

Overall, the prediction flow consists of three stages. (1) The case is sent to the LK module, if the FOL rule set is capable of producing a prediction, the result is generated. (2) The case is sent to the IB module. The case distributor classifies the case into two categories, “simple” or “complex“, based on its probability distribution. If the case is classified as simple, it is directed to the interpretable model; for cases deemed complex, it is directed to the black-box model. (3) In the final step, the selected model will make the prediction.

3.1 Legal knowledge (LK) module

One of the key innovations in our work is the integration of legal knowledge into the LK module. Legal professionals possess extensive legal knowledge, accumulated through their experience in dealing with diverse cases. For instance, a suspect with a stable residence and employment, but charged with a minor offense, may be paroled by a prosecutor.

In order to represent such legal knowledge to the machine, we utilize the First-Order Logic (FOL) rule. We leverage the power of conditional statements in FOL to capture crucial legal knowledge necessary for accurate predictions. The conditional statement is denoted as $X \rightarrow Y$, where X represents the precondition and Y represents the consequent. Preconditions can be expressed as conjunctions or disjunctions of variables, such as $X_1 \wedge X_2 \wedge \ldots \wedge X_N \rightarrow Y$. All the preconditions in these FOL rules are the fundamental elements abstracted from relevant law articles.

According to Article 81 of the Criminal Procedure Law of The People’s Republic of China, the suspect shall be detained under these three circumstances: (1) there is evidence to prove the facts of a crime and a suspect, (2) the suspect may be sentenced to imprisonment or a heavier punishment, (3) and the suspect has social harm. If the suspect does not pose a great danger to society, then he will be paroled. Since it’s costly to design FOL rules for all the cases, we chose the two most frequent charges as an attempt: murder and drunk driving, and then formulated the legal knowledge as the following FOL rules K1 and K2:

$$K_1 : \neg X_{REP} \wedge X_{JOB} \wedge X_{DRINK} \wedge X_{DRIVE} \rightarrow \neg Y$$

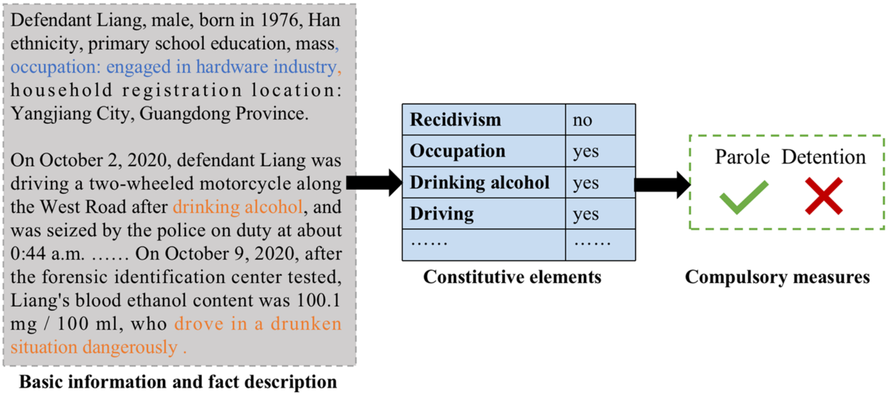
where X_{REP} is a variable representing if the suspect is a repeat offender, X_{JOB} means if the suspect has a stable job, X_{DRINK} and X_{DRIVE} represent if he has drinking or driving behavior respectively.

$$K_2 : X_{MUR} \wedge \left(X_{INJ} \vee X_{REP} \vee \neg X_{FIX}\right) \rightarrow Y$$

where X_{MUR} is a variable representing if the suspect has committed murder, X_{INJ} means if the victim was injured, and X_{FIX} represents if he has a fixed home.

As shown in Fig. 3, we utilize designed regular expressions to identify the relevant constitutive elements in the documents. If the cases satisfy the conditions specified in the FOL rules, their compulsory measure prediction results can be deduced directly, and the interpretability can be achieved naturally.

However, though the LK module is based on legal knowledge and the output can be interpreted by the FOL rules, it’s notable that the prediction is not always precise



Defendant Liang, male, born in 1976, Han ethnicity, primary school education, mass, occupation: engaged in hardware industry, household registration location: Yangjiang City, Guangdong Province.

On October 2, 2020, defendant Liang was driving a two-wheeled motorcycle along the West Road after drinking alcohol, and was seized by the police on duty at about 0:44 a.m. …… On October 9, 2020, after the forensic identification center tested, Liang's blood ethanol content was 100.1 mg / 100 ml, who drove in a drunken situation dangerously .

Recidivism	no
Occupation	yes
Drinking alcohol	yes
Driving	yes
……	……

Fig. 3 An illustration of the LK module

even if all the preconditions are met. The prediction needs to be reviewed by the human prosecutors before the final decision. We add a more detailed discussion in Sect. 5.

3.2 Model combination (IB) module

3.2.1 Case distributor

As the objectives of interpretability and prediction accuracy are often in competition, the IB module does not strive to achieve complete transparency in model prediction. Instead, it aims to establish a collaborative approach between interpretable and black-box models, leveraging their strengths in different subsets of data. To achieve this, we design a case distributor that assists in dividing the data space, allowing the interpretable and black-box models to focus on specific subsets of data for subsequent training and inference. In this way, the interpretable and black-box models can become experts in different data spaces where they can attain various achievements.

Fig. 4 illustrates the concept of the case distributor. If we only use an interpretable model such as Logistic Regression (LR) for prediction, the red line represents the decision boundary learned by LR. Cases located away from the red line can be accurately predicted by the interpretable model, and their predictions can also be explained by the interpretable algorithm. However, the accuracy of cases that are located near the red line might not be guaranteed, as the decision boundary learned by the interpretable model may not be as accurate for complex cases. Considering the powerful learning ability of black-box models, they are more suitable for processing these cases, as they can learn a more accurate decision boundary that fits complex features. Therefore, cooperation between an interpretable model and a black-box model is needed. This collaborative approach allows us to leverage the strengths of both types of models and achieve improved accuracy and interpretability in different data subsets.

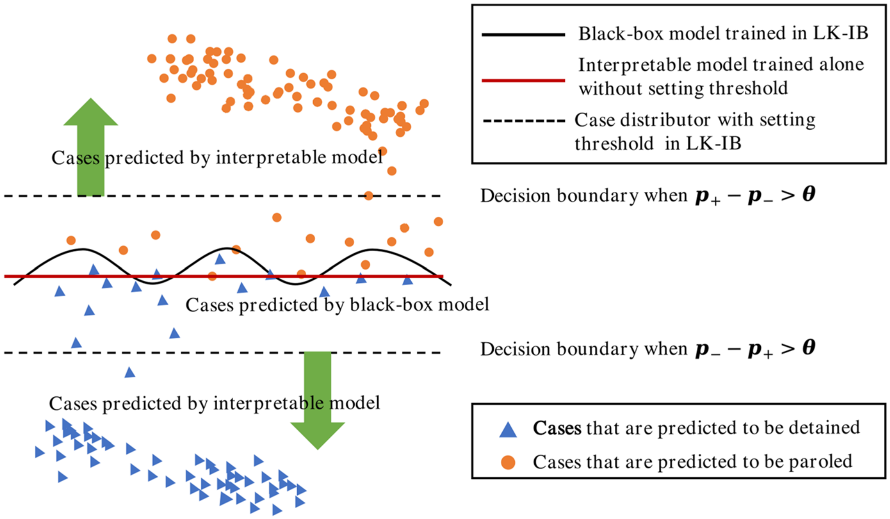

Black-box model trained in LK-IB
Interpretable model trained alone without setting threshold
Case distributor with setting threshold in LK-IB
Cases predicted by interpretable model
Decision boundary when $\boldsymbol{p}_+ - \boldsymbol{p}_- > \boldsymbol{\theta}$
Cases predicted by black-box model
Decision boundary when $\boldsymbol{p}_- - \boldsymbol{p}_+ > \boldsymbol{\theta}$
Cases predicted by interpretable model
▲ Cases that are predicted to be detained
● Cases that are predicted to be paroled

Fig. 4 The data space partition in the IB module

The case distributor calculates the probabilities of a case in different categories to determine which category it belongs to. The case distributor separates each case by taking the suspect's basic information and case description as input. If these probability values are very close, we can assume that the case is too complex to be judged by only interpretable models. So, we can divide the data space by utilizing a comprehensible model with set thresholds. The black-box model can focus on cases that are clustered badly. It's notable that interpretable models have two uses in our framework: one for case distribution and one for predicting the results. We apply two interpretable models to realize these different uses, and the former one is taken as a "case distributor" because its prediction results are not used as the final output but for distinguishing whether a case is "simple" or "complex". Then we use another interpretable model to make predictions for the simple cases. Significantly, users can select any interpretable model as the case distributor to identify simple cases. In this paper, we choose the logistic regression (LR) model as the case distributor.

Suppose $[p_+, p_-]$ represents the probability distributions over these two types of compulsory measures, which is computed by the LR model. We set a threshold value θ to divide the data space and decide which type the samples belong to. When $\|p_+ - p_-\|$ is larger than θ, the case is classified as a "simple case" and it will be sent to the interpretable model for the next processing. On the contrary, when $\|p_+ - p_-\|$ is smaller than θ, the case is classified as a "complex case" and it will be entered to the black-box model.

$$\textit{the type of case} = \begin{cases} \textit{simple} & \|p_+ - p_-\| \geq \theta \\ \textit{complex} & \|p_+ - p_-\| < \theta \end{cases} \tag{1}$$

3.2.2 Interpretable model

The choice of an interpretable model algorithm is crucial as it should be easily comprehensible for users to obtain meaningful explanations of prediction results. Therefore, we continue to employ the logistic regression algorithm as the interpretable model for predicting compulsory measures for simple cases. However, it is notable that the logistic regression model that served as the interpretable model is not the same one that served as the case distributor.

As shown in Fig. 1, the input contains a fact description and the suspect's basic information. We regard it as a word sequence, and the LR model aims to predict its corresponding compulsory measure y. Input the training dataset $D = \left\{ \left(x_1, y_1\right), \left(x_2, y_2\right), ...\left(x_n, y_n\right) \right\}$, where $x_i = \left\{ a_{i1}, a_{i2}, a_{i3}, ..., a_{ik} \right\}$ is the word sequence of the $i - th$ case,and x_i consists of k words, a_{ik} means the $k - th$ word in $i - th$ word sequence x_i. And $y_i \in \{0, 1\}$is the compulsory measure prediction result of the $i - th$ sample (0 represents that the suspect is paroled while 1 represents the detention). Features are extracted from the word sequences using TF-IDF Salton and Buckley (1988) vectorization, and suppose that $\boldsymbol{v_i}$ is the $i - th$ sample's representation. Compulsory measure probabilities are obtained as:

$$\boldsymbol{v_i} = TF - IDF(x_i) \tag{2}$$

$$P(y_i = 1 \| \boldsymbol{v_i}) = \frac{\exp\left(\boldsymbol{w^T v_i}\right)}{1 + \exp\left(\boldsymbol{w^T v_i}\right)} \tag{3}$$

$$P(y_i = 0 \| \boldsymbol{v_i}) = \frac{1}{1 + \exp\left(\boldsymbol{w^T v_i}\right)} \tag{4}$$

where $\boldsymbol{w}$ represents the weight vector in the LR model.

By Gradient Descent algorithm, we maximize the likelihood function $L(\boldsymbol{w})$ to obtain optimal parameters:

$$L(\boldsymbol{w}) = \sum_{i=1}^{n} \left[y_i \left(\boldsymbol{w^T v_i}\right) - log\left(1 + exp\left(\boldsymbol{w^T v_i}\right)\right) \right] \tag{5}$$

$$\hat{\boldsymbol{w}} = arg\max \left(\sum_{i=1}^{n} L(\boldsymbol{w}) \right) \tag{6}$$

3.2.3 Black-box model

Given that BERT Devlin et al. (2018) has significantly advanced the state of the art in various NLP tasks, including applications in the LegalAI field Chalkidis et al. (2019); Vuong et al. (2022), some researchers have adopted BERT to achieve effective performance. In our module, we also utilize the BERT model as the black-box model for predicting complex cases.

BERT utilizes transformers, which are attentive models capable of establishing relationships between words through an encoder for input and a decoder for output. In contrast to traditional NLP models that process one word at a time, BERT takes the entire sentence at once to capture the pragmatic meaning hidden between the words. In our proposed model, we do not use BERT directly, but rather a fine-tuned version in which all parameters are trained on our own dataset. The input consists of sentences containing fact descriptions and suspects' basic information. These sentences are transformed into token sequences and corresponding padding mask vectors using the bert-base-chinese pre-trained model. Subsequently, the token sequences and padding mask vectors are fed into BERT for prediction. In the notation, **h** represents the final hidden state of the input, and **W** represents the trainable parameters during model training. A pooled module and a linear module are added on top of BERT to perform binary classification:

$$P(y||\mathbf{h}) = sigmoid(W\mathbf{h}) \tag{7}$$

The loss function is a cross-entropy loss:

$$Loss = -(y \log(p) + (1-y)\log(1-p)) \tag{8}$$

3.2.4 Training strategy

Algorithm 1 Training strategy in the IB module

```
Require: $D((\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), ..., (\boldsymbol{x_n}, y_n))$
training a case distributor on the whole $D$
while X in $D$ do
    if $(\boldsymbol{x_i}, y_i)$ cannot be inferred by FOL then
        input $(\boldsymbol{x_i}, y_i)$ to case distributor
        if $||p_{i+} - p_{i-}|| > \theta$ then
            add $(\boldsymbol{x_i}, y_i)$ to the complex case subset $D_{complex}$
        else
            add $(\boldsymbol{x_i}, y_i)$ to the simple case subset $D_{simple}$
        end if
    end if
end while
training Interpretable Model on $D_{simple}$ refer to formula (6)
training Black-box Model on $D_{complex}$ refer to formula (8)
```

The training strategy is illustrated in Algorithm 1. Firstly, all the training data is used to train a case distributor. Secondly, we use first-order rules to filter out cases whose prediction results can be inferred directly from legal knowledge. The remaining cases in the training dataset are then inputted into the case distributor, which

Table 1 Statistics of dataset

#Training set	110000
#Test set	18744
Avg. #tokens in basic information	163
Avg. #tokens in fact description	168
Detention rate	41.80%

outputs the probability distribution on compulsory measures. Cases classified as "simple cases" are collected into a simple case subset for training the interpretable model in the IB module. Meanwhile, cases identified as "complex cases" by the case distributor are collected into a complex case subset for training the black-box model. Finally, the interpretable model and the black-box model complete their training process on their respective training sets.

4 Experiment

In this section, we will provide a detailed description of the dataset that was constructed for our experiment, along with the necessary parameters used in our models. We will then compare the performance of our framework with baseline methods and conduct an analysis of the effects of each module in our framework.

4.1 Dataset

The cases in our dataset were sourced from the official website of the Supreme People's Procurator of China.

We process all samples in our dataset following these steps: (1) we filter out the documents that are void (e.g, basic information is missing). (2) we remove sensitive information, such as names and other personally identifiable information, from the dataset to protect privacy. (3) we objectively split the remaining documents into three parts (e.g., basic information, fact description and compulsory measures) through the keywords. After these processing steps, we obtain a total of 119,744 available items. In our experiments, all the baselines used in this paper are trained and tested on the same dataset. The training set consists of 110,000 samples, while the testing set contains 18,744 samples. On average, each case in the dataset contains 168 sentences in the fact description and 163 sentences in the basic information, as shown in Table 1.

The dataset is mainly from four representative provinces, namely Beijing, Guangdong, Zhejiang, and Yunnan, to represent the eastern, western, southern, and northern regions of China, respectively. Our dataset reflects the overall situation in China, as these regions have distinctive geographical, economic, and political characteristics. For example, Zhejiang Province is at the forefront of judicial reform, and Guangdong Province has the highest GDP, while Yunnan Province has relatively lower economic development. Beijing is the nation's capital. The varying

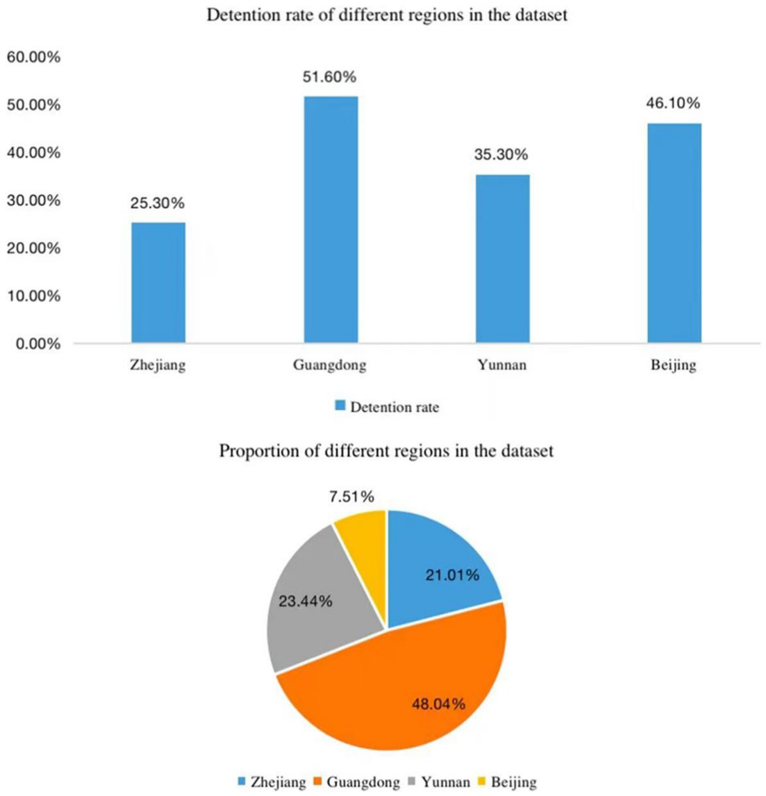


Fig. 5 The detention rate and proportion of four regions

development situations of these regions lead to significant differences in detention rates of 25.3%, 51.6%, 35.3% and 46.1%, respectively, as shown in Fig. 5. Also, data from different provinces account for different percentages in our dataset, as shown in Fig. 5.

4.2 Metrics and baselines

4.2.1 Evaluation of prediction performance

The performance of the compulsory measure prediction task is measured by classification accuracy(ACC), precision(P), recall(R), and F1(F1). Considering that both the precision and recall are calculated according to a certain category which only represents a local effect, we average them at a macro level to evaluate the performance of our model in a global aspect. The calculation is based on four indicators: False Positive (FP) stands for the number of instances that are labeled as positive while they are negative; Accordingly, False Negative (FN) is the number of instances that are labeled as negative while they are positive. True Positive (TP) and

True Negative (TN) represent the number of positive and negative instances that are correctly labeled, respectively. We adopt accuracy, precision, recall, and macro-F1 as our evaluation metrics, all the formulas are defined as follows:

$$
\begin{aligned}
&Accuracy = \frac{TP + TN}{TP + TN + FP + FN}\\
&Precision = \frac{TP}{TP + FP}\\
&Recall = \frac{TP}{TP + FN}\\
&F1 = \frac{2 * Precision * Recall}{Precision + Recall}
\end{aligned}
$$

4.2.2 Evaluation of case distributor

To evaluate the effectiveness of our case diversion approach, we employed an expert scoring method. We used our case distributor in LK-IB to generate simple and complex case datasets. From each dataset, we randomly selected 200 cases to form a new evaluation dataset. We then invited three experts in the field of law to read the fact descriptions and suspects' basic information of these cases. They evaluated the difficulty level of decision-making for each case on a scale of 0–10 and made judgments on the final compulsory measures decision. If the manual evaluation results are consistent with the final results obtained by our framework, while ensuring a certain accuracy of compulsory measure prediction, it indicates that the case diversion procedure is relatively effective.

4.2.3 Baselines

In order to evaluate the prediction performance and interpretability of our LK-IB framework, we implemented several baselines to compare these two aspects.

- **Naive Bayes** Rish et al. (2001): The naive Bayes classifier greatly simplifies learning by assuming that features are independent given class. We implement a Naive Bayes text classifier with word-level TF-IDF features.
- **Decision Tree** Safavian and Landgrebe (1991): The most important feature of Decision Tree classifiers is their capability to break down a complex decision-making process into a collection of simpler decisions, thus providing a solution that is often easier to interpret.
- **KNN** Peterson (2009): K-nearest-neighbor (KNN) classification was developed from the need to perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to determine.
- **Logistic Regression** MN and Basheer (2003): Logistic regression is a widely used statistical modeling technique with high interpretability, and we implement a Logistic Regression classifier with word-level TF-IDF features.

- **FastText** Joulin et al. (2017): FastText is defined as a light weight model, which classifies by simply adding the word vectors as the text feature.
- **Att-BiLSTM** Zhou et al. (2016): Att-BiLSTM (Attention-Based Bidirectional Long Short-Term Memory Networks) utilizes BiLSTM with an attention mechanism, which can automatically focus on the words that have a decisive effect on classification, to capture the most important semantic information in a sentence.
- **DPCNN** Johnson and Zhang (2017): DPCNN (Deep Pyramid Convolutional Neural Networks) is a wide and effective Convolutional Neural Network for deep text classification at the word-level. It is mainly composed of the region embedding layers and convolution blocks.
- **BERT** Devlin et al. (2018): BERT (Bidirectional Encoder Representations from Transformers) is a transformers model pre-trained on a large corpus of data in a self-supervised fashion. And we also conducted ablation studies with the settings as follows:
- **LK-IB w/o FOL**: w/o FOL means we remove the FOL rules, and only keep the model combination of the interpretable model and the black-box model with a case distributor.
- **LK-IB with different threshold**: we set different thresholds to the case distributor, named LK-IB0.999, LK-IB0.99, and LK-IB0.9, which represent the threshold of 0.999, 0.99, and 0.9, respectively.

4.3 Experimental details

To handle the case documents written in Chinese without spaces between words, we utilize the JIEBA tool for Chinese word segmentation. After segmentation, we apply the TF-IDF algorithm to map the segmented words into vector matrices for the case distributor and the interpretable model. Specifically, we use the TfidfVectorizer from the sklearn library to implement this approach.

In our experiments, we use BERT pre-trained models as black-box models, and the input texts are segmented into character units. We set the maximum sentence length to 256 characters, and any excess characters are removed. For training, we use the Adam optimizer Kingma and Ba (2014) with a learning rate of $3e^{-5}$, and we set the batch size to 16 for all black-box models. We train each black-box model for 3 epochs. In order to evaluate the prediction performance and interpretability of our LK-IB framework, we implemented several baselines. The parameter settings for all interpretable and black box models in the baseline model are consistent with the above.

4.4 Experimental results

4.4.1 Accuracy of compulsory measure prediction

For the CMP task, Table 2 summarizes the overall performance of our LK-IB framework and other baselines on our dataset. We also examine the effectiveness of

Table 2 The compulsory measures prediction results on our dataset. The bold values represent the best performance exhibited by all models in the experiment, i.e. the maximum values of ACC, P, R, and F1 across all models

		ACC (%)	P (%)	R (%)	F1 (%)
Interpretable models	Naive bayes	87.02	86.75	87.85	86.91
	Decision tree	94.26	94.17	93.97	94.07
	KNN	85.44	85.35	84.41	84.79
	Logistic regression	95.61	95.36	95.61	95.48
Black-box models	FastText	97.66	97.56	97.62	97.59
	Att-BiLSTM	97.87	97.77	97.82	97.80
	DPCNN	97.77	97.66	97.74	97.70
	BERT	98.00	**98.00**	97.87	97.93
Hybrid framework	LK-IB	**98.09**	**98.00**	**98.06**	**98.03**
	LK-IB w/o FOL	96.52	96.24	95.68	95.95

different components and investigate the effect of different thresholds set by the case distributor on the accuracy rate through ablation experiments. Since some machine learning models are characterized by greater simplicity and comprehensibility, they are considered as intrinsically interpretable models, including Naive Bayes, Decision Tree, KNN, and Logistic Regression. On the contrary, most neural network models show excellent predictive performance but are also characterized by greater complexity, and therefore are identified as black-box models, including FastText, Att-BiLSTM, DPCNN, and BERT. Overall, the prediction performance of LK-B whose threshold achieves the best performance on our dataset as indicated by all of the evaluation metrics. Specifically, we have several observations:

- Our proposed LK-IB framework achieves the best performance on the task of compulsory measures prediction, with an accuracy of 98.09%, which is an improvement of 7.51% on average compared to the four interpretable models, and 0.265% higher on average compared to the four black-box models. The results for other evaluation metrics, such as precision, recall, and F1-score, also demonstrate the effectiveness of our proposed framework.
- It is evident that black-box models exhibit superior predictive performance compared to interpretable models, despite their lack of interpretability. Additionally, BERT stands out as the best-performing model among all the baselines, with the highest precision at 98.00% and an accuracy that is only slightly lower than our framework. Among the interpretable models, Logistic Regression demonstrates better performance compared to Naive Bayes, Decision Tree, and KNN. In the following ablation studies, we adapt both BERT and Logistic Regression as the black-box and interpretable prediction models, respectively, in our LK-IB framework.

All of the above experiments demonstrate that the injection of legal knowledge benefits our framework in terms of both interpretability and prediction accuracy.

Table 3 The expert scores for evaluating case diversion effect

	Difficult Level	
	Simple cases subset	Complex cases subset
Expert 1	1.98	8.45
Expert 2	2.14	7.42
Expert 3	2.02	8.23
Average	2.05	8.03

Furthermore, the combination of IM and BM results in more explainable predictions without significantly sacrificing accuracy.

4.4.2 Evaluation of case distributor

Table 3 shows the expert scores of simple and complex cases as judged by our framework, based on the ratings provided by three legal experts. The average score for simple cases is 2.05, while the average score for complex cases is 8.03. This indicates that our case diversion approach is effective in distinguishing between simple and complex cases, because its classification results align with the experts' cognitive understanding in the legal domain.

4.4.3 Ablation study

We conducted ablation experiments to demonstrate the effectiveness of FOL and hybrid models respectively in our proposed LK-IB framework.

First, as shown in Table 2, the accuracy of LK-IB is 1.57% higher than LK-IB w/o FOL. The only variable in this comparison is whether FOL is introduced. Consequently, the result indicates the effectiveness of FOL for case diversion and inference.

Second, as shown in Table 2, the accuracy of LK-IB w/o FOL is 96.52%, which is 1.48% lower than that of the single BERT model. This observation confirms that the design of hybrid models can improve interpretability without sacrificing too much accuracy, making it a promising approach for achieving a balance between accuracy and interpretability in CMP task.

Finally, we intend to explore the impact of different thresholds on LK-IB's predictive performance. We set the diversion thresholds of LK-IB at 0.999,0.99 and 0.9 respectively. When the threshold of the case distributor is set differently while keeping the total training sample size fixed, the division of the training set will also vary. A higher threshold value means a higher standard for a sample to be classified as a simple case. As a result, the training dataset for the black-box model (BM) in LK-IB will be larger, and the training dataset for the interpretable model (IM) in LK-IB will be smaller, as shown in Table 4. We can see that the accuracy of LK-IB whose threshold is 0.999 is the highest, the predictive accuracy improves as the diversion

Table 4 The division of training data under different thresholds

The threshold value (d)	Size of simple case set	Size of complex case set	ACC	Size of cases predicted by LK module	Total size
0.9	57914	13371	97.59%	38715	1100000
0.99	32547	38738	98.03%		
0.999	19403	51882	98.09%		

threshold gets higher. On the other hand, once the threshold setting is higher, the number of simple cases classified by case distributor will decrease, which means the number of explainable cases will decrease.

4.4.4 Impact of hyper-parameters

Our framework has an important hyper-parameter: the set threshold in the case distributor. Here, we study the impact of this hyper-parameter on the performance of our model. Following the approach used in Wang and Lin (2021), we adopt the definition that the transparency of a hybrid model $f =< f_l,f_b >$ on D_{total} is the percentage of data processed by f_l (where D_{total} represents the total number of data samples in the dataset, and f_l and f_b represent the interpretable model and the black-box model, respectively). It's worth mentioning that in our framework, there are some samples that are inferred by First-Order Logic (FOL), and FOL can also be viewed as an interpretable model. Hence, f_l refers to both FOL and the logistic regression (LR) model in our framework. The transparency is computed by:

$$transparency = \frac{D_{FOL} + D_{LR}}{D_{total}}$$

where D_{FOL} means the sample subset inferred by FOL and D_{LR} represents the sample subset predicted by LR model.

Figure 6 shows the trade-off between transparency and predictive accuracy of all hybrid models in our experiment. It appears that when the threshold set in the case distributor is closer to 1, the accuracy of LK-IB increases while the transparency decreases. This could be because a higher threshold results in more samples being processed by the black-box model. As a result, finding a middle ground where partial transparency and good predictive performance can coexist is possible, and users can adjust the threshold in the case distributor based on their desired level of transparency and tolerance for loss in prediction accuracy. In our experiments, we set the threshold to 0.999, as it results in the best accuracy for our framework while still maintaining an acceptable level of interpretability. This allows us to compare our framework with the baselines in a consistent manner.

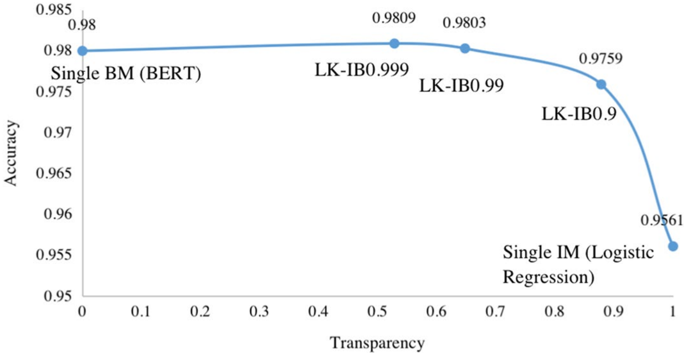


Fig. 6 The trade-off between transparency and accuracy of hybrid models. The X-aris represents the transparency, and the Y-aris represents the accuracy on the test set

Table 5 The prediction accuracy of the data from different regions

	Zhejiang	Guangdong	Beijing	Yunnan
ACC	96.94%	98.48%	96.72%	98.75%

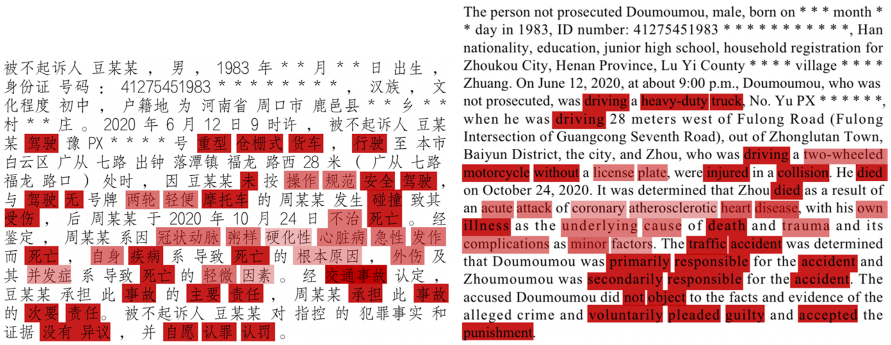
4.4.5 Experimental results on different regions

To check the suitability of the database we built for training the model, we analyzed the prediction accuracy of the data from different provinces in the test data, and the results are shown in Table 5. It can be seen that although the detention rate varies from province to province, the accuracy rate reaches over 96% in different provinces, proving that our dataset can train models with high generalization ability.

4.4.6 Case study

In this section, we choose a representative example of a simple case from our experiment, in order to provide an intuitive illustration of how the interpretable model in LK-IB works for compulsory prediction. As shown in Fig. 7, the shade of color of the text directly reflects the weight value of each word trained in the interpretable model. Words with a darker background color indicate higher weights.

The content of the case demonstrates that the suspect has a regular place of residence, pleads guilty, and accepts punishment voluntarily, despite the fact that his breach of traffic laws resulted in the victim’s death. According to the law, this case does not meet the requirements for detention, and the prediction result given by the interpretable model is parole. From Fig. 7, we can observe that the interpretable model is able to capture key facts and relevant information about the suspects, which are important for determining the final compulsory measures. Based on this, we have reason to believe that the prediction results obtained by the interpretable model are explainable, as the rationale and prediction result provided by the interpretable model are consistent with real legal scenarios.



被不起诉人 豆某某 ， 男 ， 1983 年 * * 月 * * 日 出生 ， 身份证 号码 ： 41275451983 * * * * * * * * ， 汉族 ， 文化程度 初中 ， 户籍地 为 河南省 周口市 鹿邑县 * * 乡 * * 村 * * 庄 。 2020 年 6 月 12 日 9 时许 ， 被不起诉人 豆某某 驾驶 豫 PX * * * * 号 重型 仓栅式 货车 ， 行驶 至 本市 白云区 广从 七路 出钟 落潭镇 福龙 路西 28 米 （ 广从 七路 福龙 路口 ） 处时 ， 因 豆某某 未 按 操作 规范 安全 驾驶 ， 与 驾驶 无 号牌 两轮 轻便 摩托车 的 周某某 发生 碰撞 致其 受伤 ， 后 周某某 于 2020 年 10 月 24 日 不治 死亡 。 经 鉴定 ， 周某某 系因 冠状动脉 粥样 硬化性 心脏病 急性 发作 而 死亡 ， 自身 疾病 系 导致 死亡 的 根本原因 ， 外伤 及其 并发症 系 导致 死亡 的 轻微 因素 。 经 交通事故 认定 ， 豆某某 承担 此 事故 的 主要 责任 ， 周某某 承担 此 事故 的 次要 责任。 被不起诉人 豆某某 对 指控 的 犯罪事实 和 证据 没有 异议 ， 并 自愿 认罪 认罚 。

The person not prosecuted Doumoumou, male, born on * * * month * * day in 1983, ID number: 41275451983 * * * * * * * * * * *, Han nationality, education, junior high school, household registration for Zhoukou City, Henan Province, Lu Yi County * * * * village * * * * Zhuang. On June 12, 2020, at about 9:00 p.m., Doumoumou, who was not prosecuted, was driving a heavy-duty truck, No. Yu PX * * * * * *, when he was driving 28 meters west of Fulong Road (Fulong Intersection of Guangcong Seventh Road), out of Zhonglutan Town, Baiyun District, the city, and Zhou, who was driving a two-wheeled motorcycle without a license plate, were injured in a collision. He died on October 24, 2020. It was determined that Zhou died as a result of an acute attack of coronary atherosclerotic heart disease, with his own illness as the underlying cause of death and trauma and its complications as minor factors. The traffic accident was determined that Doumoumou was primarily responsible for the accident and Zhoumoumou was secondarily responsible for the accident. The accused Doumoumou did not object to the facts and evidence of the alleged crime and voluntarily pleaded guilty and accepted the punishment.

Fig. 7 Visualization of words' weight in a simple case, which can be viewed as the explanation of the interpretable model

5 Ethical discussion

Because compulsory measures are closely tied to the basic human rights of suspects, any subtle miscalculation may trigger serious consequences. For example, the FOL rules are not precise for all situations, even if all the preconditions are met, a suspect still has the possibility to be paroled by the prosecutors in the real world. Thus, ethical issues require further investigation.

First of all, we want to emphasize that the aim of the proposed models is not to replace humans but to inform and enhance the decision-making processes of prosecutors. The role of prosecutors in ensuring fairness and justice remains critical, as they will review the results generated by the algorithm as a final safeguard. Thus, the potential bias in the model (e.g., job bias and residence bias) can be finally addressed by human prosecutors. In fact, many models proposed in the field of LegalAI prior to this were aimed at providing a reference for humans rather than replacing human judgment Zhou et al. (2022); Bi et al. (2022).

In the future, we will strive to train more impartial computational models by removing potentially discriminatory data items from the training data and replacing them with neutral alternatives Bolukbasi et al. (2016), or by utilizing causal inference mechanisms Wu et al. (2020) to identify and mitigate confounding variables. These measures will reduce the potential biases in the model.

6 Conclusion

In this paper, we investigate the compulsory measure prediction (CMP) task in China from the perspective of interpretability. We propose a novel hybrid framework named LK-IB for the CMP task, which leverages legal knowledge and model combination. Specifically, we first translate the legal knowledge into the first-order logic(FOL) rules. Then, if the compulsory measure can't be predicted by the FOL rule, LK-IB will use a case distributor to categorize whether the case is "simple" or “complex”. Finally, an interpretable model will predict the results for the simple cases and a black-box model will predict the results for the complex case.

We collect legal documents from the website of the Supreme People’s Prosecutor of China and construct a novel dataset to conduct the experiment. Our experimental results show that LK-IB improves the interpretability of prediction results without sacrificing too much accuracy. Ultimately, our framework has practical usefulness in real legal scenarios.

In future research, we aim to improve our proposed framework in four ways. (1) First, we will comprehensively consider making case diversion more precise by carefully selecting case distributors and investigating how different types of distributors affect case prediction results. We plan to design more detailed legal knowledge graphs to optimize case diversion and make it more appropriate for real-world judicial scenarios. (2) Second, we will explore how to enhance the interpretability of complex cases by utilizing tools such as attention mechanisms Du et al. (2019) and causal inference Kuang et al. (2020). This will enable us to provide more transparent and understandable explanations for the predictions made by our model. (3) Third, we propose to extend the case diversion idea presented in this paper to a wider range of judicial scenarios, including civil case judgments, which constitute a significant portion of simple cases. Moreover, our approach is not limited to judicial prediction tasks; it can also be applied to court’s view generation Wu et al. (2020) for simple cases, while leaving trial reasons of complex cases to judges or more sophisticated computational models, thereby significantly enhancing trial efficiency. (4) Fourth, we will focus on improving our database by incorporating more complex cases, such as those involving multiple suspects and the combined use of different coercive measures. These cases present greater challenges in terms of predictive accuracy and interpretability. Besides, we will remove potentially discriminatory data items from the training data and replace them with neutral alternatives, and then strive to train more impartial computational models. These improvements will further enhance the effectiveness, interpretability, and applicability of our framework in real-world legal scenarios.

Appendix A: Relevant Law Articles

Here we provide with some relevant law articles in Criminal Procedure Law of the People’s Republic of China, which the prosecutor must following when making a decision on compulsory measures.

- **Article 67** A people’s court, a people’s prosecutor, and a public security authority may grant bail to a suspect or defendant under any of the following circumstances: (1) the suspect or defendant may be sentenced to supervision without incarceration, limited incarceration, or an accessory penalty only; (2) the suspect or defendant may be sentenced to fixed-term imprisonment or a heavier penalty but will not cause danger to the society if granted bail; (3) the suspect or defendant suffers a serious illness, cannot take care of himself or herself or is a pregnant woman or a woman who is breastfeeding her own baby, and will not cause danger to the society if granted bail; or (4) The term of custody of the suspect or defendant has expired but the case has not been closed, and a bail is necessary. Bail shall be executed by a public security authority.

- **Article 72** The authority deciding on a bail shall decide the amount of a bond after fully considering the need to ensure normal legal proceedings, the danger of the person to be bailed to the society, the nature and circumstances of the case, the gravity of the possible punishment, the financial condition of the person to be bailed, and other factors.
- **Article 80** The arrest of a suspect or defendant must be subject to the approval of a people's prosecutor or a decision of a people's court and be executed by a public security authority.
- **Article 81** Where there is evidence to prove the facts of a crime and a suspect or defendant may be sentenced to imprisonment or a heavier punishment, if residential confinement is insufficient to prevent any of the following dangers to society, the suspect or defendant shall be arrested: (1) the suspect or defendant may commit a new crime; (2) there is an actual danger to national security, public security, or social order; (3) the suspect or defendant may destroy or forge evidence, interfere with the testimony of a witness, or make a false confession in collusion; (4) the suspect or defendant may retaliate against a victim, informant, or accuser; or (5) the suspect or defendant attempts to commit suicide or escape. In the process of approving or deciding an arrest, the nature and circumstances of the suspected crime, the admission of guilt, and the acceptance of punishment, among others, of a suspect or defendant shall be considered as factors of a possible danger to the society. Where there is evidence to prove the facts of a crime and a suspect or defendant may be sentenced to fixed-term imprisonment of 10 years or a heavier punishment or there is evidence to prove the facts of a crime and a suspect or defendant who once committed an intentional crime or has not been identified may be sentenced to imprisonment or a heavier punishment, the suspect or defendant shall be arrested. Where a suspect or defendant waiting for trial on bail or under residential confinement seriously violates the provisions on bail or residential confinement, the suspect or defendant may be arrested.

Funding The research leading to these results received funding from the National Key Research and Development Program of China No. 2021YFC3300300 and the National Social Science Foundation under Grant Agreement No.21CFX068.

Declarations

Conflicts of interest The authors have no relevant financial or non-financial interests to disclose.

References

Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE Access 6:52138–52160

Adler P, Falk C, Friedler SA, Rybeck G, Scheidegger C, Smith B, Venkatasubramanian S (2016) Auditing black-box models by obscuring features. arXiv preprint arXiv:1602.07043

Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R et al (2020) Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fus 58:82–115

Bao Q, Zan H, Gong P, Chen J, Xiao Y (2019) Charge prediction with legal attention. In: CCF International Conference on Natural Language Processing and Chinese Computing, pp. 447–458. Springer

Bi S, Zhou Z, Pan L, Qi G (2022) Judicial knowledge-enhanced magnitude-aware reasoning for numerical legal judgment prediction. Artif Intell Law 1–34

Bolukbasi T, Chang K-W, Zou JY, Saligrama V, Kalai AT (2016) Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Adv Neural Inf Process Syst 29

Branting LK, Pfeifer C, Brown B, Ferro L, Aberdeen J, Weiss B, Pfaff M, Liao B (2021) Scalable and explainable legal prediction. Artif Intell Law 29(2):213–238

Brennan T, Dieterich W, Ehret B (2009) Evaluating the predictive validity of the compas risk and needs assessment system. Crim Justice Behav 36(1):21–40

Chalkidis I, Androutsopoulos I, Aletras N (2019) Neural legal judgment prediction in english. arXiv preprint arXiv:1906.02059

Chugh N (2021) Risk assessment tools on trial: Lessons learned for "ethical ai" in the criminal justice system. In: 2021 IEEE International Symposium on Technology and Society (ISTAS), pp. 1–5. https://doi.org/10.1109/ISTAS52410.2021.9629143

Cohen TH, Lowenkamp C (2018) Revalidation of the federal pretrial risk assessment instrument (ptra): Testing the ptra for predictive biases. Available at SSRN

Desmarais SL, Zottola SA, Duhart Clarke SE, Lowder EM (2021) Predictive validity of pretrial risk assessments: a systematic review of the literature. Crim Justice Behav 48(4):398–420

Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Dieterich W (2010) Kent County Pretrial Services outcomes study: developing and testing the COMPAS pretrial release risk scale. Northpointe

Dionne G (2013) Risk management: history, definition, and critique. Risk Manag Insur Rev 16(2):147–166

Do P-K, Nguyen H-T, Tran C-X, Nguyen M-T, Nguyen M-L (2017) Legal question answering using ranking svm and deep convolutional neural network. arXiv preprint arXiv:1703.05320

Du M, Liu N, Hu X (2019) Techniques for interpretable machine learning. Commun ACM 63(1):68–77

Fawei B, Pan JZ, Kollingbaum M, Wyner AZ (2019) A semi-automated ontology construction for legal question answering. New Gener Comput 37(4):453–478

Gan L, Kuang K, Yang Y, Wu F (2021) Judgment prediction via injecting legal knowledge into neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 12866–12874

Hacker P, Krestel R, Grundmann S, Naumann F (2020) Explainable AI under contract and tort law: legal incentives and technical challenges. Artif Intell Law 28(4):415–439

Jiang X, Ye H, Luo Z, Chao W, Ma W (2018) Interpretable rationale augmented charge prediction system. In: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pp. 146–151

Johnson R, Zhang T (2017) Deep pyramid convolutional neural networks for text categorization. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 562–570. Association for Computational Linguistics, Vancouver, Canada . https://doi.org/10.18653/v1/P17-1052. https://aclanthology.org/P17-1052

Joulin A, Grave E, Bojanowski P, Mikolov T (2017) Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp. 427–431. Association for Computational Linguistics, Valencia, Spain. https://aclanthology.org/E17-2068

Kim M-Y, Goebel R (2017) Two-step cascaded textual entailment for legal bar exam question answering. In: Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law, pp. 283–290

Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980

Kuang K, Li L, Geng Z, Xu L, Zhang K, Liao B, Huang H, Ding P, Miao W, Jiang Z (2020) Causal inference. Engineering 6(3):253–263

Li Q, Zhang Q (2021) Court opinion generation from case fact description with legal basis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 14840–14848

Liu Y-H, Chen Y-L (2018) A two-phase sentiment analysis approach for judgement prediction. J Inf Sci 44(5):594–607

Liu L, Zhang W, Liu J, Shi W, Huang Y (2021) Interpretable charge prediction for legal cases based on interdependent legal information. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE

Long S, Tu C, Liu Z, Sun M (2019) Automatic judgment prediction via legal reading comprehension. In: China National Conference on Chinese Computational Linguistics, pp. 558–572 . Springer

Luo B, Feng Y, Xu J, Zhang X, Zhao D (2017) Learning to predict charges for criminal cases with legal basis. arXiv preprint arXiv:1707.09168

Madras D, Pitassi T, Zemel R (2018) Predict responsibly: improving fairness and accuracy by learning to defer. Adv Neural Inf Process Syst 31

Madsen A, Reddy S, Chandar S (2021) Post-hoc interpretability for neural nlp: a survey. ACM Comput Surv (CSUR)

Miron M, Tolan S, Gómez E, Castillo C (2021) Evaluating causes of algorithmic bias in juvenile criminal recidivism. Artif Intell Law 29(2):111–147

Mn H, Basheer I (2003) Comparison of logistic regression and neural network-based classifiers for bacterial growth. Food Microbiol 20:43–55. https://doi.org/10.1016/S0740-0020(02)00104-1

Peeters R, Schuilenburg M (2018) Machine justice: Governing security through the bureaucracy of algorithms. Inf Polity 23(3):267–280

Peterson LE (2009) K-nearest neighbor. Scholarpedia 4(2):1883

Rish I *et al.* (2001) An empirical study of the naive bayes classifier. In: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, vol. 3, pp. 41–46

Safavian SR, Landgrebe D (1991) A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 21(3):660–674

Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Inf Proces Manag 24(5):513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Singh A, Mohapatra S (2021) Development of risk assessment framework for first time offenders using ensemble learning. IEEE Access 9:135024–135033

Vuong YT-H, Bui Q.M, Nguyen H-T, Nguyen T-T-T, Tran V, Phan X-H, Satoh K, Nguyen L-M (2022) Sm-bert-cr: a deep learning approach for case law retrieval with supporting model. Artif Intell Law, 1–28

Wang T, Lin Q (2021) Hybrid predictive models: when an interpretable model collaborates with a black-box model. J Mach Learn Res 22:137

Wu Y, Kuang K, Zhang Y, Liu X, Sun C, Xiao J, Zhuang Y, Si L, Wu F (2020) De-biased court's view generation with causality. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 763–780

Xu Z, Li X, Li Y, Wang Z, Fanxu Y, Lai X (2020) Multi-task legal judgement prediction combining a subtask of the seriousness of charges. In: China National Conference on Chinese Computational Linguistics, pp. 415–429 . Springer

Xu N, Wang P, Chen L, Pan L, Wang X, Zhao J (2020) Distinguish confusing law articles for legal judgment prediction. arXiv preprint arXiv:2004.02557

Yang H, Deng W, Wang G, Wang F, Li S (2020) Interpretable legal judgment prediction based on improved conditional classification tree. In: Developments of Artificial Intelligence Technologies in Computation and Robotics: Proceedings of the 14th International FLINS Conference (FLINS 2020), pp. 336–343. World Scientific

Ye H, Jiang X, Luo Z, Chao W (2018) Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. arXiv preprint arXiv:1802.08504

Zhong H, Xiao C, Tu C, Zhang T, Liu Z, Sun M (2020) How does nlp benefit legal system: a summary of legal artificial intelligence. arXiv preprint arXiv:2004.12158

Zhou S, Liu Y, Wu Y, Kuang K, Zheng C, Wu F (2022) Similar case based prison term prediction. In: Artificial Intelligence: Second CAAI International Conference, CICAI 2022, Beijing, China, August 27–28, 2022, Revised Selected Papers, Part III, pp. 284–297. Springer

Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, Xu B (2016) Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (volume 2: Short Papers), pp. 207–212

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Xiang Zhou[1] · Qi Liu[1] · Yiquan Wu[2] · Qiangchao Chen[1] · Kun Kuang[2]

✉ Yiquan Wu
wuyiquan@zju.edu.cn

✉ Kun Kuang
kunkuang@zju.edu.cn

Xiang Zhou
0020355@zju.edu.cn

Qi Liu
liuqi_jx@zju.edu.cn

Qiangchao Chen
22102078@zju.edu.cn

[1] School of Guanghua Law, Zhejiang University, Hangzhou, China

[2] School of Computer Science and Technology, Zhejiang University, Hangzhou, China