Journal of Network Intelligence
Taiwan Ubiquitous Information

Volume **11**, Number **1**, February **2026**

Automatic Classification of Legal Documents Assisted by Image Recognition Technology

Cong Wen[1,*]

[1]Institute of Marxism,
Henan Vocational College of Nursing, Anyang 455000, P. R. China
wen_cong161@126.com

Shan-Shan Liu[2]

[2]School of technology,
Asia Pacific University of Technology & Innovation, Kuala Lumpur 56000, Malaysia
523272231@qq.com

[*]Corresponding author: Cong Wen

Received November 11, 2024, revised February 14, 2025, accepted June 3, 2025.

Abstract. *In the digital age, the proliferation of legal documents has posed unprecedented challenges for document management and classification systems. Traditional manual methods are not only labor-intensive but also struggle with the growing volume and complexity of data, necessitating the development of automated classification techniques. This paper introduces an innovative automated framework for the classification of legal documents using image recognition technology. The framework incorporates Optical Character Recognition (OCR), N-Gram tokenization, TF-IDF feature extraction, and an XGBoost classifier to achieve high accuracy and efficiency in document categorization. The OCR component converts image files into text, which is then processed by the N-Gram tokenization to capture contextual information. TF-IDF feature extraction assesses the significance of vocabulary, forming feature vectors for the text. The XGBoost classifier then uses these vectors to classify the documents. Extensive experiments demonstrate that our approach surpasses conventional methods in both classification accuracy and operational efficiency. The study also provides an in-depth analysis of the contributions of each model component and suggests future research directions, such as expanding support for multi-language documents, enhancing performance on small sample datasets, integrating advanced deep learning techniques, and testing the model's generalization capabilities across various legal domains. This research significantly advances the automation of legal document classification and contributes to the digital transformation of the legal sector.*
Keywords: image recognition techniques; N-Gram segmentation; TF-IDF; XGBoost

1. **Introduction.** As we enter the digital era, there has been a significant surge in the volume of legal documents [1]. This exponential growth presents novel challenges for the management and categorization of such documents. Manual classification of legal documents, the traditional method, is increasingly impractical due to its extensive demands on time and labor. Moreover, as data volumes expand, this approach encounters significant difficulties in maintaining both precision and efficiency. With the onset of the digital era, there has been a significant surge in the volume of legal documents, presenting new challenges for document management and categorization, as an effective means of extracting text from images, has been widely used in a variety of fields, including legal, medical, and

financial. However, the application of OCR technology only solves the problem of text extraction, and how to further classify the extracted text accurately and efficiently is the focus of current research [2].

This study proposes an automatic classification method for legal documents based on image recognition technology that combines Optical Character Recognition (OCR), N-Gram Segmentation, TF-IDF feature extraction and XGBoost classifier. Firstly, legal documents in image format are converted to text data using OCR technique. Then, the contextual information in the text is captured by N-Gram participle and the importance of words is evaluated using the TF-IDF method to achieve the feature vector representation of the text. Ultimately, the extracted features are subjected to classification utilizing the XGBoost algorithm.

This study not only improves the automation level of legal document classification, but also provides a new direction for future research on related technologies. By optimising the OCR technique, improving the N-Gram disambiguation and TF-IDF feature extraction methods, and customising and optimising the XGBoost classifier, our model shows substantial improvements across various evaluation metrics. This not only advances the development of automatic legal document classification technology, but also supports the digital transformation of the legal industry.

1.1. **Related work.** In recent years, image recognition and text classification techniques have been widely used in the field of automatic classification of legal documents. Sun et al. [3] studied the utilization of Optical Character Recognition (OCR) technology facilitates the digital conversion of extensive document collections, especially in legal and historical documents, and proposed a method to improve the recognition rate of OCR. His research provides an important technical basis for digitisation of legal documents. Sil and Roy [4] proposed a machine learning based classification method by analysing structured and unstructured data in legal documents. The method combines text extraction and classification techniques and significantly improves the classification accuracy of legal documents in practical applications. Their research utilised OCR techniques to extract text from images and used Support Vector Machines (SVMs) to classify them, and the experimental results proved that the classification results were better than traditional manual classification methods.

In the field of deep learning, Huang et al. [5] proposed a legal document classification method based on Named Entity Recognition (NER), which relies on domain adaptation of rule annotators to automatically extract and annotate important entity information in legal documents. This study provides a new idea based on semantic analysis for the classification of legal documents, but the performance of this method is more limited when facing multi-language and multi-format documents.

In terms of hybrid models, Omurca et al. [6] proposed a document classification system combining OCR, TF-IDF and XGBoost. The system first extracts textual information from documents by OCR technique, then uses TF-IDF for feature extraction and finally classifies them by XGBoost. The trials have demonstrated that employing the XGBoost model yields superior precision and expedience when managing voluminous and intricate legal documents, outperforming conventional classification techniques.

Although these researches provide an important foundation for automatic classification of legal documents, current approaches exhibit certain constraints when it comes to handling the intricacies and variety present in sophisticated document types, especially in dealing with scenarios with diverse text formats and complex legal terminology [7].

Therefore, how to further improve the classification accuracy and optimise the processing efficiency remains an important challenge in the current research on automatic legal document classification.

1.2. **Contribution.** Within the scope of this manuscript, we introduce a pioneering framework for the automated categorization of legal documents. This framework integrates a suite of advanced methodologies designed to bolster the precision and expedience of the classification process. The framework's implementation involves an initial preprocessing stage, where noise removal, binarization, and skew correction are applied to improve OCR accuracy.

Our design firstly utilises optimised OCR techniques for pre-processing and text extraction of legal document images, taking measures such as noise removal, binarisation and skew correction, especially for the complex typography and possible low resolution of legal documents. Next, we implemented the N-Gram segmentation approach to extract contextual cues within the text. This method is especially crucial for precisely identifying technical jargon present in legal documents. We also improved the TF-IDF feature extraction method by adjusting the calculation of word frequency and inverse document frequency to make the model more sensitive to key terms in legal texts.

In addition, we fine-tuned the XGBoost classifier by optimizing hyperparameters—including the learning rate, tree depth, and number of trees—to better adapt it for the complexities of classifying legal documents. Our experiments aimed to assess not just accuracy, but also a set of evaluation metrics—including precision, recall, and F1 score—to thoroughly examine the model's performance. Using ablation studies, we meticulously assessed the impact of every model component, providing insights into the significance and function of various techniques within the context of legal document classification. Finally, although our design is specific to legal documents, its core techniques and framework have cross-domain applicability and can be generalised to other types of professional document classification tasks. These innovations not only promote the development of automatic legal document classification technology, but also provide new directions and ideas for future research on related technologies. We believe that these innovations will help improve the automation of legal document processing and support the digital transformation of the legal industry.

2. Theoretical analysis.

2.1. **Image recognition technology.** Image recognition technology is crucial for extracting text from images, with applications in various fields including legal document management [8]. It finds broad applications across various industries, including autonomous vehicles, surveillance systems, and medical imaging analysis. On the basis of image recognition, the system extracts useful information by analysing the features of the image [9]. Generally speaking, the fundamental steps in image recognition comprise image preprocessing, feature extraction, and the classification phase. Within this process, feature extraction frequently leverages deep learning architectures like Convolutional Neural Networks (CNNs), which are adept at identifying complex patterns and high-level features within images [10]. The underlying principle of CNNs can be formulated as:

$$y = f(W * X + b) \tag{1}$$

where X denotes the input image, W signifies the convolutional filters, b represents the bias, and y is the resultant output following the application of the activation function. Through multilayer convolutional operations, complex patterns and features in the image are gradually extracted and used for subsequent classification tasks.

Optical Character Recognition (OCR) is an important application area of image recognition technology, specialising in the extraction of text from images or scans [11]. The core objective of OCR is to recognise and convert characters in an image into machine-readable text. This approach is particularly essential for handling extensive legal document datasets, as it enables swift identification and retrieval of crucial data from the documents. The working principle of OCR can be broken down into two main steps: text area detection and character recognition [12].

In the phase of text region identification, the algorithm employs image processing methods, such as edge detection and morphological operations, to pinpoint the textual segments within the image. This process can be encapsulated by the following formula:

$$T_d = \text{Detect}(I) \tag{2}$$

where I is the input image and T_d is the detected text region. In the character recognition stage, the system uses a classification algorithm to convert the pixel data in the text region into characters. Assuming that the detected text region is represented by a pixel matrix X, the character recognition process can be achieved by the classification function $f(x)$:

$$C = f(X) \tag{3}$$

where C is the character or word recognised.

In the automatic classification of legal documents, OCR is a key step in extracting text data, but to improve the accuracy of classification, it is also necessary to combine advanced data mining with focusing models [13]. The text data extracted by OCR can be further processed by Natural Language Processing (NLP) technology to extract key information such as case number, law articles, etc [14]. Leveraging these extracted features, the system can then deploy targeted models, like Random Forest or Support Vector Machine, to effectuate precise document classification.

The core idea of the focusing model is to perform efficient classification based on the extracted features. Assuming that the text data obtained from OCR is T, the focusing model extracts the features F based on this data, and then classifies the document by a classifier:

$$y = \text{Classify}(F) \tag{4}$$

where $F = \text{Extract Features}(T)$ and y is the classification result of the document. By integrating with the focal model, OCR enhances the precision and expedites the process of automated document classification.

2.2. **Focused model.** With OCR technology converting legal documents into textual data, the next challenge we face is to extract key information from these texts that can significantly impact classification decisions [15]. The text focus model employs the N-Gram disambiguation technique to segment the text into finer units, thereby capturing the contextual nuances within the textual content. Then, the model utilizes the TF-IDF (Term Frequency-Inverse Document Frequency) approach to assess the significance of each word within the document. This process results in a feature vector representation for the text, as depicted in Figure 1.

(1) N-Gram Segmentation. The N-Gram participle technique, also known as the n-tuple model, is a statistically based method for processing natural language [16]. It treats words, phrases or words in a text as random variables that follow a certain probability distribution. The essence of the N-Gram model is that the likelihood of a word occurring is influenced not merely by the word itself, but also by the context provided by the preceding words. This model posits that the emergence of a word at the Nth position is solely contingent upon the $(N-1)$ words that come before it [17]. Consequently, the

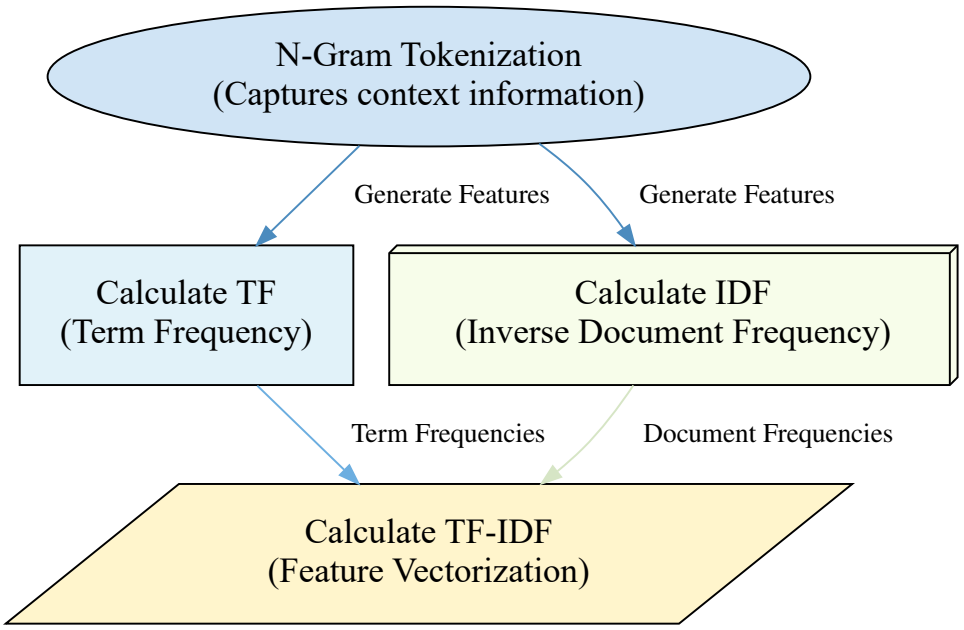

Figure 1. A Focused model

probability of an entire sentence occurring can be viewed as the multiplication of the individual conditional probabilities starting from the first word up to the Nth word.

The sequence of an N-Gram model can be visualized as a chain of elements: $NGRAM(1), NGRAM(2)$ with each $NGRAM(i)$ signifying a consecutive string of n words, commencing from the ith word.

$$P(W_1, W_2, \ldots, W_n) = \prod_{i=1}^{n} P(W_i | W_1, \ldots, W_{i-1}) \tag{5}$$

Although the N-Gram model can provide powerful text representation, it also faces the problem that the parameter space grows exponentially with the increase of N, which may lead to data sparsity and large consumption of computational resources [18]. To address these issues, smaller values of N, such as 2 (Bi-Gram) or 3 (Tri-Gram), are usually chosen in practical applications to strike a balance between the model's complexity and the practicality of computation.

In legal document classification, the N-Gram segmentation technique can help us capture the contextual relationship between words and provide richer feature information for the classification model. For example, the Bi-Gram model can split the phrase "legal document classification" into two consecutive lexical units, "legal document" and "document classification", thus providing useful information for classification. The phrase "legal document classification" can be split into two consecutive lexical units, "legal document" and "document classification", thus providing useful information for classification.

To deal with data sparsity, techniques such as Laplace smoothing are often used [19]. Laplace smoothing modifies the probability estimates by incrementing the counts with a small constant—typically 1—to mitigate the issue of zero probabilities. This adjustment can be represented by the equation:

$$P(W_{i+1} | W_i) = \frac{\text{count}(W_i, W_{i+1}) + 1}{\text{count}(W_i) + N} \tag{6}$$

where $\text{count}(W_i, W_{i+1})$ denotes the frequency with which the word pair (W_i, W_{i+1}) occurs in the dataset, while N represents the total count of unique word pairs across the dataset.

In this way, the N-Gram disambiguation technique not only provides fine-grained features of the text, but also helps us to construct a classification model for legal documents that is both accurate and efficient.

(2) TF-IDF vectorisation: TF-IDF is calculated by the product of term frequency and inverse document frequency, if the word T_i has a high number of occurrences in a certain text (TF high) and rarely occurs in other texts (IDF high), then T_i is considered to have a high role in category differentiation (TF-IDF high) [20]. It can be seen that TF portrays the importance of T_i for a particular text and IDF portrays the importance of T_i for the whole text set (corpus). The steps for its computation are as follows.

Calculate TF:

$$\mathrm{TF}_{i,j} = \frac{N_{i,j}}{\sum_k N_{k,j}} \tag{7}$$

where $N_{i,j}$ represents the frequency of term T_i in document D_j, while the denominator denotes the total count of all terms present in document D_j. Here it is standardised by dividing by the denominator, and other standardisation methods can be used in practice.

Calculate the IDF:

$$\mathrm{IDF}_i = \log \frac{|D|}{|\{j : T_i \in D_j\}|} \tag{8}$$

where $|\{j : T_i \in D_j\}|$ denotes the count of documents that include the term T_i; $|D|$ signifies the overall count of documents in the dataset.

IDF is also calculated in various ways, e.g.:

$$\mathrm{IDF}_i = \log \left(\frac{|D|}{|\{j : T_i \in D_j\}| + 1} \right) \tag{9}$$

Here, the denominator +1 is to eliminate the case where the denominator is 0.

Calculate the TF-IDF:

$$\text{TFI-DF} = \mathrm{TF} \times \mathrm{IDF} \tag{10}$$

2.3. **XGBoost classification model.** After obtaining the text features, XGBoost is used to classify the text [21]. The idea of XGBoost algorithm is the same as the idea of boosting tree in integrated learning. Boosting tree is the learning framework in integrated learning, and its training is based on the residuals. The model at stage n takes as input the residuals from the prediction outcomes of the preceding model at stage $n-1$, with the training of each model occurring sequentially over time [22]. The goal of each model training is to make the residuals smaller and smaller, and it can be seen that the final training result is actually a look-add of the training result of each model, as in Figure 2.

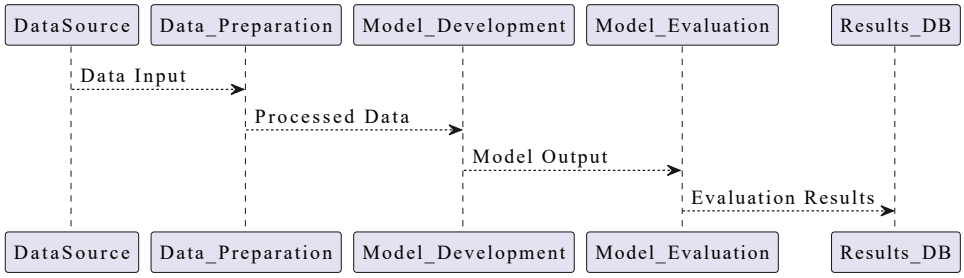

Figure 2. A XGBoost model

The models in the aforementioned training process are identified as "trees", and each cycle of XGBoost training introduces a new tree to address the discrepancies between the actual and forecasted values from the prior cycle, thus gradually approximating the actual values [23].

The objective function O_{bj} during XGBoost training model is:

$$O_{bj} = L + \sum_{i=1}^{n} \Omega(f_i) \tag{11}$$

where L is the error term and $\sum_{i=1}^{t} \Omega(f_i)$ is the complexity function term.

$$L = \sum_{i} (y_i - \hat{y}_i)^2 \tag{12}$$

$$\hat{y}_i = \sum_{k=1}^{k} f_k(x_i), f_k \in F \tag{13}$$

where x_i represents the training dataset; y_i denotes the label associated with x_i; F signifies the entire feature space; f is an individual feature within F; K indicates the total count of trees; and $\hat{y}_i$ is the predicted value for the training instance x_i after all trees have been applied.

The loss term L is calculated by aggregating the discrepancies between the predicted values following the training on the entire dataset and the actual values. Concurrently, the predictions generated in each iteration serve as adjustments to the residuals from the preceding iteration [24]. Suppose the forecast generated by the model at step s is denoted as $\hat{y}_i^{(s)}$, then the forecast at each iteration can be articulated as:

$$\hat{y}_i^{(0)} = 0 \tag{14}$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \tag{15}$$

$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \tag{16}$$

$$\hat{y}_i^{(3)} = f_1(x_i) + f_2(x_i) + f_3(x_i) = \hat{y}_i^{(2)} + f_3(x_i) \tag{17}$$

...

$$\hat{y}_i^{(s)} = \sum_{k=1}^{s} f_k(x_i) = \hat{y}_i^{(s-1)} + f_s(x_i) \tag{18}$$

The predicted value $\hat{y}_i^{(s)}$ at step s is derived from the sum of the prediction $f_s(x_i)$ made at step s and the estimate from the previous step $s-1$. This aggregated result is then plugged into the error term to formulate the objective function.

$$O_{bj}^{(s)} = \sum_{i=1}^{n} (y_i - (\hat{y}_i^{(s-1)} + f_s(x_i)))^2 + \sum_{i=1}^{n} \Omega(f_i) \tag{19}$$

In the context of the s-th training iteration, the outcomes from step s and earlier are considered known and can be regarded as constants for subsequent derivations. The complexity function term $\Omega(f_s)$ is primarily influenced by the number of leaf nodes T in the tree and the weights w_j associated with these leaf nodes, which can be expressed as:

$$\Omega(f_s) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} W_j^2 \tag{20}$$

where γ, λ are hyperparameters that regulate the number of leaf nodes and the weights, respectively.

3. **A model for automatic classification of legal documents assisted by image recognition techniques.**

3.1. **Modeling framework.** Our end-to-end automated process is robust against variations in image quality, our advanced preprocessing techniques that cater to images with low resolution or complex typography. One of the core innovations of our modelling framework is the implementation of an end-to-end automated process that converts image inputs of legal documents directly into classification outputs without any human intervention. This process covers the key steps of image preprocessing, text extraction, feature extraction and classification decision making, ensuring efficient and highly accurate processing, see Figure 3.

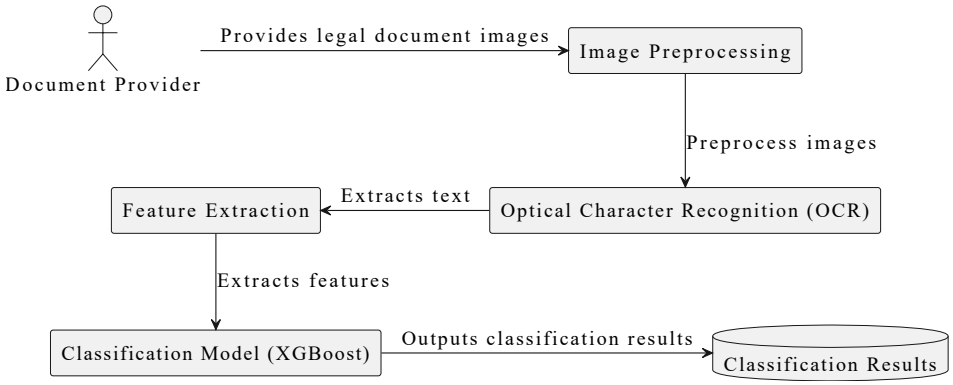

Figure 3. A mixed model for automatic classification of legal documents assisted

(1) Image pre-processing. Image preprocessing is a crucial step in the process of automatic classification of legal documents, which directly affects the quality of subsequent text extraction and the accuracy of classification. The preprocessing process mainly includes operations such as de-noising, binarisation and skew correction, aiming to improve image quality and provide clear image data for text extraction and feature extraction.

De-noising is the first step in pre-processing, which aims to remove random noise from the image and make the text clearer. We usually use a Gaussian filter to smooth the image and reduce the effect of noise on subsequent processing.

This is followed by the binarisation step, which converts the image to contain only black and white, making the text more clearly separated from the background. This step is achieved by setting a threshold T with the following formula:

$$I_{binarized}(x,y) = \begin{cases} 255 & \text{if } I(x,y) > T \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

In this way, the text portion of the image is converted to white (or black) while the non-text portion is converted to black (or white), providing a clear boundary for text extraction.

Finally, the tilt correction step is used to adjust the image tilt due to incorrect shooting or scanning angle. This step corrects the image by detecting the tilt angle of the text lines and applying a rotation transformation to ensure that the text lines are horizontally aligned. The tilt angle θ can be calculated using the following formula:

$$\theta = \arctan\left(\frac{\sum_{(x,y)\in\text{text region}} y}{\sum_{(x,y)\in\text{text region}} x}\right) \tag{22}$$

This formula determines the tilt angle by calculating the centre-of-mass offset of the text region, and then applies a rotation transformation to correct the image.

With these preprocessing steps, we are able to significantly improve the recognition accuracy of our OCR technique and provide high-quality image data for subsequent feature extraction and classification decisions. The combination of these techniques ensures that our model is able to handle legal document images of various qualities, including those of low quality due to poor scanning or shooting conditions, thus laying a solid foundation for efficient and accurate automatic classification of legal documents.

(2) Text Extraction. In the legal document classification process, documents are input in the form of images, which first need to be converted into processable text data by OCR technology. We used Tesseract OCR, which has good performance in multilingual environments and is able to handle legal documents with low resolution or complex typography. In the image pre-processing stage, we performed noise removal, binarisation and skew correction on the documents to improve the recognition accuracy of OCR.

Text extracted by OCR is cleaned and pre-processed, including the removal of stop words, special characters and symbols to ensure the effectiveness of subsequent processing. This step is especially important for documents in different formats (e.g. contracts, judgements, pleadings, etc.), as legal documents often have complex typography and terminology.

(3) Feature extraction. In the process of text feature extraction, this paper adopts the combination of N-Gram segmentation and TF-IDF to enhance the accuracy of text representation. N-Gram segmentation is an effective method of text feature representation, which is able to capture the contextual relationship between words, especially in legal documents, where there are often phrases or terms that have special meanings.

Following the extraction of N-Gram features, we proceed to convert the text into vectors utilizing the TF-IDF method. This approach not only assesses the significance of a word within an individual document but also diminishes the impact of words that frequently appear across the entire document collection through the IDF component. This is particularly beneficial for capturing specialized legal terminology present in legal documents.

(4) Classifier design and optimisation. After feature extraction, we use XGBoost as a classifier. XGBoost is a high-performance machine learning algorithm that builds on the Gradient Boosting Decision Tree (GBDT) framework. It is adept at handling high-dimensional and sparse data sets, and it incorporates self-regularization during the training phase, which helps to mitigate overfitting. The core of XGBoost lies in the reduction of error and minimisation of the loss function by adding weak classifiers step by step.

To enhance the model's performance even further, we employed cross-validation and grid search techniques to fine-tune the hyperparameters of the XGBoost algorithm. This optimization targeted aspects such as the depth of the trees, the learning rate, and the quantity of weak classifiers utilized in the training process [25]. The specific hyperparameters that were optimized encompass:

- learning rate: modulates the impact of each individual tree's prediction on the overall loss.
- max_depth: restricts the depth of the trees to avoid overfitting.
- subsample: prevents overfitting of the model to certain training samples.

The optimised XGBoost classifier shows good generalisation performance on multiple types of legal documents.

Through this end-to-end automated process, our model not only improves the speed and accuracy of legal document classification, but also reduces the possibility of human error. The implementation of this automated process is a major innovation in the field of legal document classification, which provides a new solution for automated processing

of legal documents. We believe that this automated process will further promote the automation and intelligent development of legal document processing technology.

3.2. **Evaluation indicators.** Evaluation metrics such as precision rate, accuracy rate, recall rate, and F1 value are mainly used to evaluate the effectiveness of bidding document classification achieved in this paper, in which the F1 value is a comprehensive assessment of the recall rate and precision rate [26]. In multi-classification problems, the final results of F1, recall and other metrics are obtained after macro-averaging, i.e., the values of F1 and other metrics are calculated according to each class (label) and then averaged to obtain the results. The following is an example of a 3-classification problem to introduce the specific calculation method. The confusion matrix of the 3-classification problem is shown in Table 1.

Table 1. A confusion matrix for evaluating classification models.

Category	The number predicted to be Class A	The number predicted to be Class B	The number predicted to be Class C
A	AA	AB	AC
B	BA	BB	BC
C	CA	CB	CC

In the case of category A, this can be calculated by using the confusion matrix in Table 3:

$$TP_A = AA \tag{23}$$

$$FP_A = BA + CA \tag{24}$$

$$FN_A = AB + AC \tag{25}$$

Where TP_A denotes the count of instances belonging to class A that are accurately identified as such; FP_A represents the count of instances not belonging to class A but erroneously classified as class A; and FN_A indicates the count of instances that are part of class A yet incorrectly assigned to other classes.

$$p_A = \frac{TP_A}{TP_A + FP_A} \tag{26}$$

$$r_A = \frac{TP_A}{TP_A + FN_A} \tag{27}$$

$$F1_A = \frac{2 \times p_A \times r_A}{p_A + r_A} \tag{28}$$

where p_A is the precision rate of Class A samples; r_A is the recall rate of Class A samples; $F1_A$ is the F1 score of Class A samples.

Using the same method the precision p_B, recall r_B, and F1 score $F1_B$ can be calculated for the class B samples and the precision p_C, recall r_C, and F1 score $F1_C$ can be calculated for the class C samples, so the final evaluation result of the model is:

By employing the same approach, the precision p_B, recall r_B, and F1 score $F1_B$ can be determined for class B instances, and similarly, the precision p_C, recall r_C and F1 score $F1_C$ can be ascertained for class C instances. Consequently, the overall model evaluation is derived from these metrics.

$$p = \frac{p_A + p_B + p_C}{3} \tag{29}$$

$$r = \frac{r_A + r_B + r_C}{3} \tag{30}$$

$$F1 = \frac{F1_A + F1_B + F1_C}{3} \tag{31}$$

The overall accuracy of the model α is:

$$\alpha = \frac{AA + BB + CC}{AA + AB + AC + BA + BB + BC + CA + CB + CC} \tag{32}$$

where F1 is the overall F1 score of the model [27]. The accuracy rate α indicates the general performance of the predictions, signifying the proportion of samples that are correctly identified. The precision rate p denotes the model's ability to accurately identify negative samples, with higher values indicating greater discriminative power. The recall rate r measures the model's effectiveness in recognizing positive samples, where a higher value suggests a stronger ability to detect positive instances. The F1 score serves as a comprehensive assessment, striking a balance between precision and recall.

4. **Performance testing and analysis.**

4.1. **Ablation Experiments.** We first constructed a base model that only utilises OCR technology for text extraction and does not involve feature extraction or classification processes. This stage aims to assess the fundamental performance of OCR in recognizing text and to establish a baseline for the integration of additional components later on.

On the basis of the base model, we introduce the N-Gram segmentation technique. By generating N-Gram feature representations, the model successfully identifies the contextual interconnections among words within the textual content. Experimental results show that the introduction of N-Gram disambiguation significantly improves the classification accuracy of the model, which suggests that N-Gram disambiguation plays an important role in recognising phrases and specific terms in legal texts.

Next, we added TF-IDF feature extraction to the model. TF-IDF further enhances the quality of the feature representation by assessing the significance of words within the document, thus improving the classification results. By comparing the models using only N-Gram disambiguation and using both N-Gram disambiguation and TF-IDF features, we find that the introduction of TF-IDF significantly improves the classification performance. TF-IDF makes the model more focused on the key legal terms by reducing the impact of frequently occurring words.

Finally, to evaluate the performance of the classifier, We conducted a comparison between XGBoost and several other prevalent classification methods, including Support Vector Machine (SVM) and Random Forest. The data presented in Table 2 illustrates that XGBoost excels in terms of both accuracy and F1 scores, markedly surpassing other classification models. This suggests that XGBoost's attributes are particularly advantageous for managing the complexities of legal text classification challenges.

The ablation experiments results indicate that the inclusion of N-Gram segmentation and TF-IDF feature extraction significantly enhances the model's ability to capture contextual relationships and specialized legal terminology.

4.2. **Comprehensive experiments.** The extensive experimental section is designed to thoroughly assess the efficacy of our proposed automatic legal document classification model when applied to real-world scenarios and to contrast its performance with other cutting-edge classification techniques. We've chosen a range of established classification techniques—including SVMs, Random Forests, and Naive Bayes—as well as advanced deep learning models like CNNs and LSTM networks—to benchmark against our approach

Table 2. Results of ablation experiments

Model	**Accuracy**	**Precision**	**Recall**	**F1**
Baseline Models	65.3	63.1	60.5	61.8
N-Gram Segmentation	74.1	72.5	70.3	71.4
N-Gram + TF-IDF	80.6	78.3	76.8	77.5
SVM	78.2	76.1	74.0	75.0
Random Forest	76.5	74.2	72.1	73.1
XGBoost	84.5	82.0	80.0	80.9

for classifying legal documents. This comparison allows us to evaluate the performance of each method within this specific domain.

The dataset used for the experiments covers a wide range of types of legal documents, including contracts, judgements and pleadings. The dataset was partitioned, allocating 70% for training, 15% for validation, and the remaining 15% for testing purposes. To maintain impartiality, each model undergoes training on an identical dataset and subsequent assessment on a consistent test dataset. We employ a suite of evaluation metrics—comprising accuracy, precision, recall, and the F1 score—to provide a comprehensive analysis of the performance across all models.

Throughout the training phase, we conducted hyper-parameter tuning for both the SVM and Random Forest classifiers to achieve optimal results for the given tasks. For SVMs, we experimented with both linear and RBF kernels, while for Random Forests, we optimized the number and depth of the trees using grid search techniques. In the case of deep learning models, we meticulously adjusted their architectural configurations and hyperparameters—such as the learning rate, batch size, and epochs—to enhance the classification outcomes for both CNNs and LSTMs.

The experimental data is depicted in Figure 4, illustrating a comparison of the performance across various models on the test dataset. A comparative analysis of our model with other state-of-the-art classification techniques, such as SVMs and Random Forests, reveals that XGBoost outperforms these methods in terms of accuracy and F1 score, particularly when classifying complex and high-dimensional legal document data.

In conclusion, our extensive experiments confirm the efficacy of our proposed model and highlight the suitability of various algorithms for classifying legal documents. For future endeavors, we aim to investigate the integration of these algorithms to potentially elevate the model's performance and reliability.

5. **Conclusion.** Within this document, we introduce a framework for the automated categorization of legal documents that leverages image recognition technologies, addressing the challenges of efficiency and precision that are commonly encountered with conventional manual sorting approaches. Our method converts images of legal documents into text data through OCR technology, captures the key information of the text using N-Gram segmentation and TF-IDF feature extraction techniques, and applies the XGBoost algorithm to classify these features. In a comprehensive set of trials, we demonstrate how our approach enhances both the precision and expeditiousness of classifying legal documents, surpassing the capabilities of traditional manual methods.

Despite the positive results, our study has some limitations. Firstly, our model may face challenges in processing legal documents in multilingual and non-standard formats, as OCR techniques and feature extraction methods may need to be adapted for specific

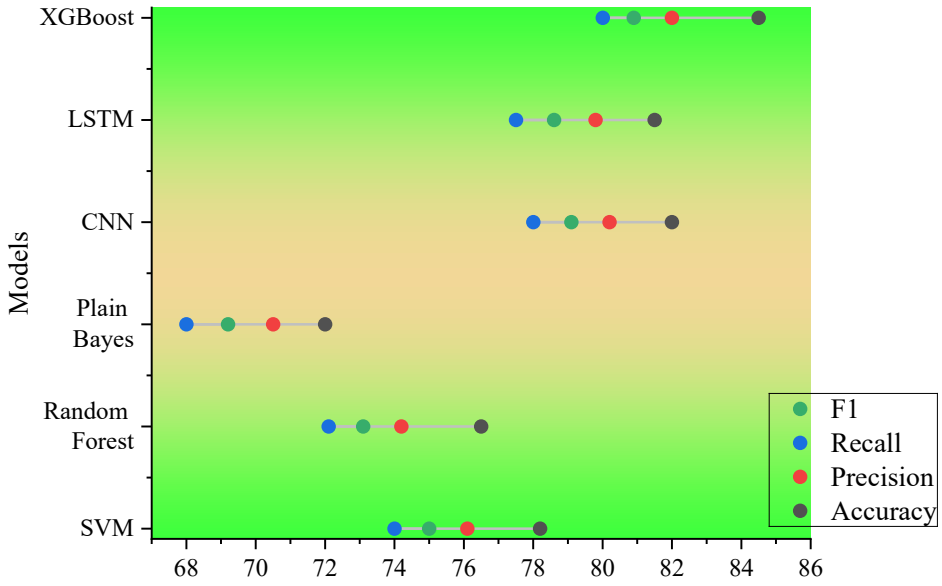

Figure 4. Synthesize the results of the experiment

languages and formats. Second, our model may not perform as well on small sample datasets as large-scale datasets, which limits its application in data-scarce domains. In addition, the generalisation ability of the model has not been fully validated in diverse legal domains, which needs to be further explored in future research.

To address these limitations, our future research directions will focus on the following areas. Firstly, we plan to extend the model to support multilingual and different formats of legal documents, which may involve the development of more advanced OCR techniques and adaptive feature extraction methods. Next, we intend to investigate the potential of semi-supervised and unsupervised learning techniques to bolster the model's performance when dealing with datasets that have limited samples. Furthermore, we plan to explore the incorporation of advanced deep learning methodologies, especially the latest developments in natural language processing, into our framework. This integration is expected to significantly enhance the precision and stability of our classification system. Finally, we plan to test and optimise our model in a wider range of legal domains and real-world application scenarios to validate its generalisation ability and usefulness.

In conclusion, the automatic classification method for legal documents based on image recognition technology proposed in this paper provides an effective solution for efficient management and automated processing of legal documents. Although there are some limitations, we believe that through continuous research and improvement, our method will make an important contribution to the digital transformation of the legal field.

REFERENCES

[1] S. Morales, K. Engan, and V. Naranjo, "Artificial intelligence in computational pathology—challenges and future directions," *Digital Signal Processing,* vol. 119, pp. 103196, 2021.

[2] S. Singh, N. K. Garg, and M. Kumar, "Feature extraction and classification techniques for handwritten Devanagari text recognition: a survey," *Multimedia Tools and Applications,* vol. 82, no. 1, pp. 747–775, 2023.

[3] W. Sun, L. M. Liu, W. Zhang, and J. C. Comfort, "Intelligent OCR processing," *Journal of the American Society for Information Science,* vol. 43, no. 6, pp. 422–431, 1992.
[4] R. Sil and A. Roy, "Machine learning approach for automated legal text classification," *International Journal of Computer Information Systems and Industrial Management Applications,* vol. 13, pp. 10–10, 2021.
[5] H. Huang, H. Wang, and D. Jin, "A Low-Cost Named Entity Recognition Research Based on Active Learning," *Scientific Programming,* vol. 2018, no. 1, p. 1890683, 2018.
[6] S. I. Omurca, E. Ekinci, S. Sevim, E. B. Edinc, S. Eken, and A. Sayar, "A document image classification system fusing deep and machine learning models," *Applied Intelligence,* vol. 53, no. 12, pp. 15295–15310, 2023.
[7] A. Roth, "Store vulnerability window (SVW): A filter and potential replacement for load re-execution," *Journal of Instruction Level Parallelism,* vol. 8, no. 1, pp. 1–22, 2006.
[8] A. Tolba, A. El-Baz, and A. El-Harby, "Face recognition: A literature review," *International Journal of Signal Processing,* vol. 2, no. 2, pp. 88–103, 2006.
[9] S. Dabbaghchian, M. P. Ghaemmaghami, and A. Aghagolzadeh, "Feature extraction using discrete cosine transform and discrimination power analysis with a face recognition technology," *Pattern Recognition,* vol. 43, no. 4, pp. 1431–1440, 2010.
[10] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial Intelligence Review,* vol. 53, pp. 5455–5516, 2020.
[11] S. Saoji, A. Eqbal, and B. Vidyapeeth, "Text recognition and detection from images using pytesseract," *Journal of Interdisciplinary Cycle Research,* vol. 13, pp. 1674–1679, 2021.
[12] Y. Alginahi, "Preprocessing techniques in character recognition," *Character Recognition,* vol. 1, pp. 1–19, 2010.
[13] C. Patel, A. Patel, and D. Patel, "Optical character recognition by open source OCR tool tesseract: A case study," *International Journal of Computer Applications,* vol. 55, no. 10, pp. 50–56, 2012.
[14] S. N. Laique, U. Hayat, S. Sarvepalli, B. Vaughn, M. Ibrahim, J. McMichael, K. N. Qaiser, C. Burke, A. Bhatt, and C. Rhodes, "Application of optical character recognition with natural language processing for large-scale quality metric data extraction in colonoscopy reports," *Gastrointestinal Endoscopy,* vol. 93, no. 3, pp. 750–757, 2021.
[15] S. R. Wong and G. Fisher, "Comparing and using occupation-focused models," *Occupational Therapy in Health Care,* vol. 29, no. 3, pp. 297–315, 2015.
[16] Y. Doval and C. Gomez-Rodriguez, "Comparing neural-and N-gram-based language models for word segmentation," *Journal of the Association for Information Science and Technology,* vol. 70, no. 2, pp. 187–197, 2019.
[17] K. Nowakowski, M. Ptaszynski, and F. Masui, "Mingmatch—a fast n-gram model for word segmentation of the ainu language," *Information,* vol. 10, no. 10, pp. 317, 2019.
[18] J. J. Bapu, D. J. Florinabel, Y. H. Robinson, E. G. Julie, R. Kumar, V. T. N. Ngoc, L. H. Son, T. M. Tuan, and C. N. Giap, "Adaptive convolutional neural network using N-gram for spatial object recognition," *Earth Science Informatics,* vol. 12, no. 4, pp. 525–540, 2019.
[19] V. Cherian and M. Bindu, "Heart disease prediction using Naive Bayes algorithm and Laplace Smoothing technique," *Int. J. Comput. Sci. Trends Technol,* vol. 5, no. 2, pp. 68–73, 2017.
[20] M. Sandhya, S. Sarika, S. Anukriti, and A. Sushila, "Automatic Text Categorization on News Articles," *International Journal of Engineering and Techniques,* vol. 2, no. 3, pp. 33–38, 2016.
[21] S. Bhattacharya, S. R. K. S, P. K. R. Maddikunta, R. Kaluri, S. Singh, T. R. Gadekallu, M. Alazab, and U. Tariq, "A novel PCA-firefly based XGBoost classification model for intrusion detection in networks using GPU," *Electronics,* vol. 9, no. 2, p. 219, 2020.
[22] S. Alagarsamy and V. James, "RNN LSTM-based deep hybrid learning model for text classification using machine learning variant xgboost," *International Journal of Performability Engineering,* vol. 18, no. 8, p. 545, 2022.
[23] S. Ghosal and A. Jain, "Depression and suicide risk detection on social media using fasttext embedding and xgboost classifier," *Procedia Computer Science,* vol. 218, pp. 1631–1639, 2023.
[24] R. A. Stein, P. A. Jaques, and J. F. Valiati, "An analysis of hierarchical text classification using word embeddings," *Information Sciences,* vol. 471, pp. 216–232, 2019.
[25] S. BN and C. Akki, "Sentiment prediction using enhanced XGBoost and tailored random forest," *International Journal of Computing and Digital Systems,* vol. 10, no. 1, pp. 199–191, 2021.
[26] B. H. Ahmed and A. S. Ghabayen, "Review rating prediction framework using deep learning," *Journal of Ambient Intelligence and Humanized Computing,* vol. 13, no. 7, pp. 3423–3432, 2022.

[27] H. Huang, H. Xu, X. Wang, and W. Silamu, "Maximum F1-score discriminative training criterion for automatic mispronunciation detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 23, no. 4, pp. 787–797, 2015.