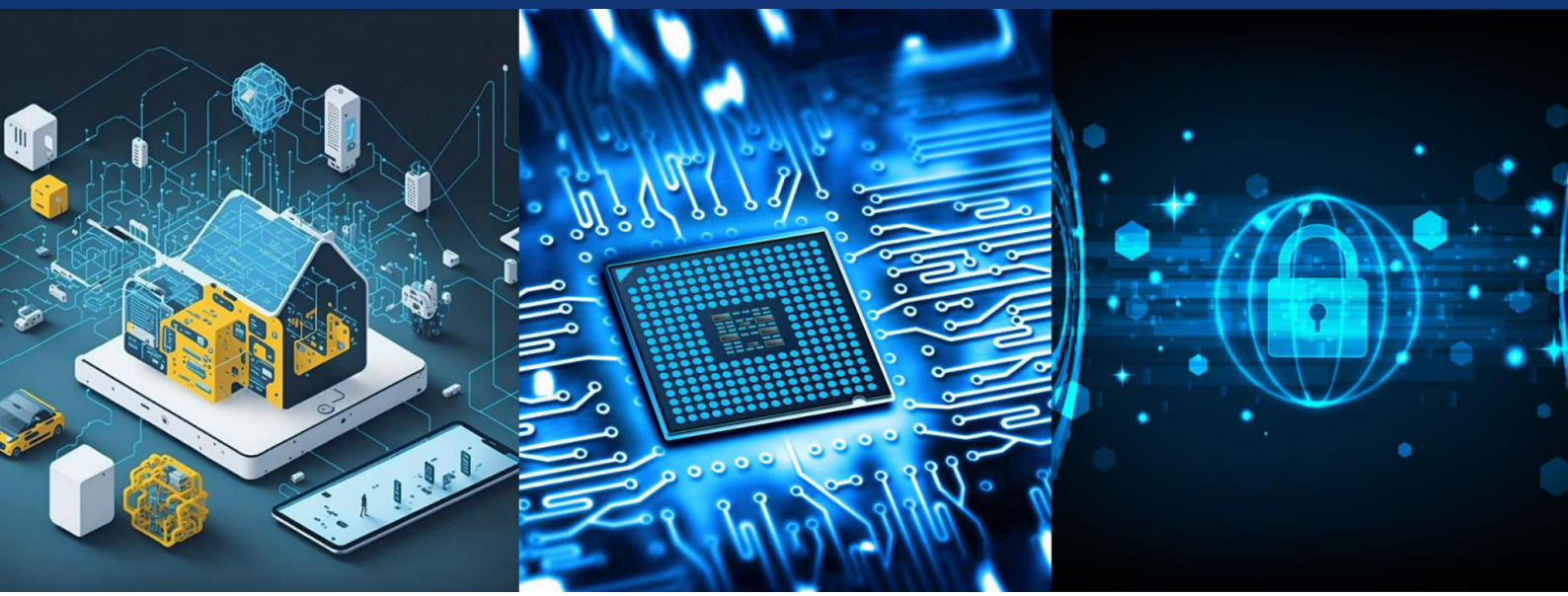
IJIRCCE

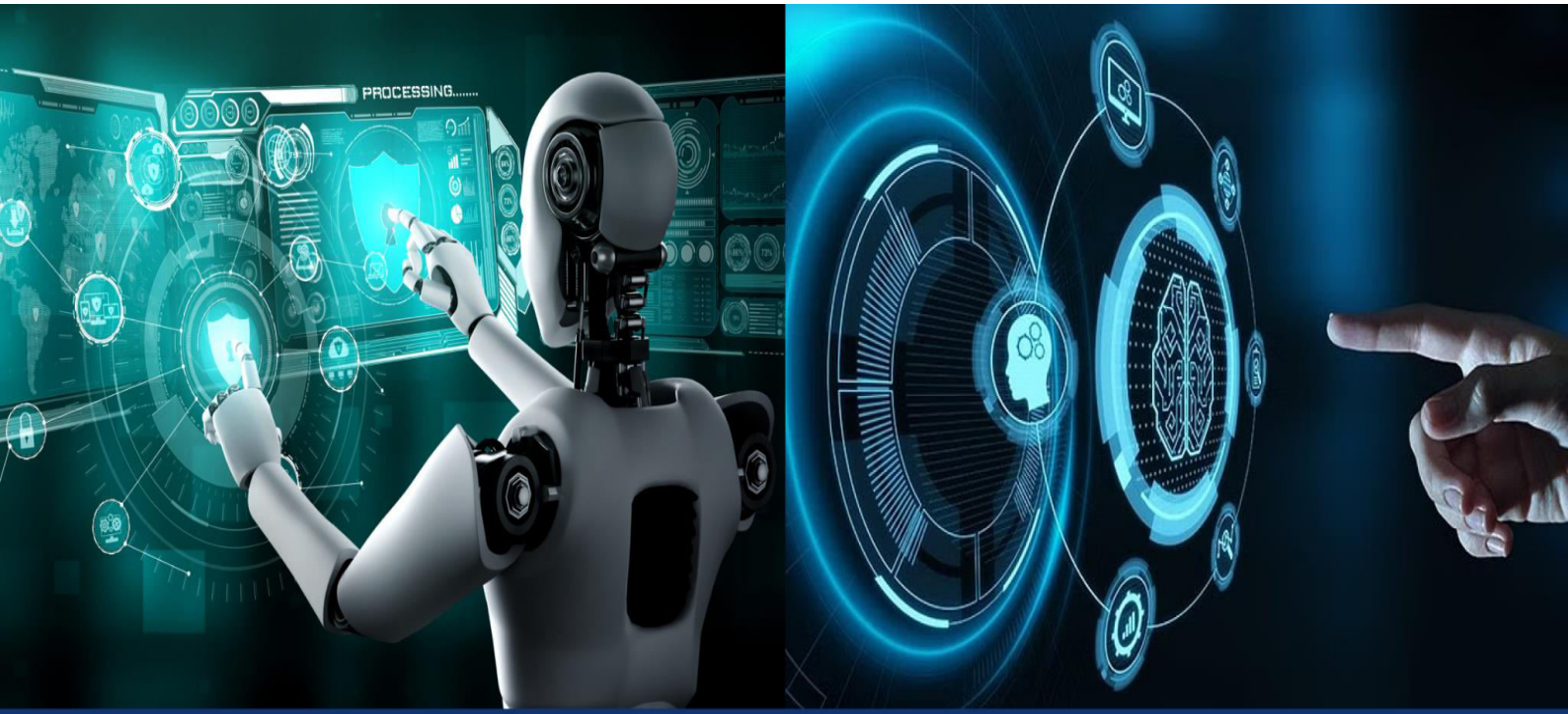
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 14, Issue 4, April 2026

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404010

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Hybrid Ensemble Learning and NLP based Risk Assessment for Legal Document Analysis

Mr. B. Avinash, K. Kundan Sai, K. Venkata Srinivasa Rao, M. Madhukar, P. Naga Vamsi

Assistant Professor, Dept. of Information Technology, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, India

B. Tech Student, Dept. of Information Technology, VVIT, Nambur, Guntur, Andhra Pradesh, India

B. Tech Student, Dept. of Information Technology, VVIT, Nambur, Guntur, Andhra Pradesh, India

B. Tech Student, Dept. of Information Technology, VVIT, Nambur, Guntur, Andhra Pradesh, India

B. Tech Student, Dept. of Information Technology, VVIT, Nambur, Guntur, Andhra Pradesh, India

ABSTRACT: The increasing adoption of digital workflows of corporate communications has produced contract volumes that routinely exceed thousands of documents per legal department annually, demanding automated and auditable review pipelines for automated legal document review. Traditional manual interrogation is resource-demanding, prone to human error, and difficult to scale to the thousands of contracts handled annually by legal departments. This paper proposes a Hybrid Deterministic and Probabilistic Risk Ensemble (HD-PRE) framework for automated legal clause risk detection. The framework synthesizes deterministic rule based Regex triggers with probabilistic ensemble classifiers Random Forest and Gradient Boosting over a fused TF-IDF and Boolean feature space. Hybrid Risk Assessor leverages the advantages of both confidence scores, which are provided by probability, and circuit breakers, which are hard coded, to provide strong legal risk pattern detection for new, unseen, or shifted legal risk patterns, even in the presence of significant dataset shift. The dataset used contains 13,000 legal clauses, with 10,000 in the train, 2,500 in Test-General, and 500 in Test-Shifted, with an overall 14.6% dataset risk ratio, drawn from CUAD and LEDGAR. Performance evaluations of the proposed system have shown better performance with 94.81% and 85.6% accuracy on Test-General and Test-Shifted, with 0.97 and 0.87 detection rates, better than other strong baselines, including Legal-BERT, which are all deep neural models.

KEYWORDS: Legal Document Analysis, Contract Risk Detection, Ensemble Learning, TF-IDF, Hybrid Machine Learning, NLP, Anomaly Detection, Dataset Shift, Random Forest, Gradient Boosting.

I. INTRODUCTION

The new legal environment is marked by the rapid rise of the data crisis. The digitalization of corporate communication has led to the unprecedented volume of contractual agreements. The corporate legal department is constantly engaged in the management of thousands of critical contracts every year. The scope of the contracts varies widely and includes Master Service Agreements, Non-Disclosure Agreements, regular employment contracts, and vendor contracts. The financial and legal implications of even one such liability clause are catastrophic. The need to review the contracts in an exhaustive and thorough manner is the prime necessity. The process of interrogating the documents in the conventional way is extremely resource demanding and is marked by difficulties in scaling the process [1], [2].

In recent years, the rapid progress in Artificial Intelligence (AI) and Natural Language Processing (NLP) has led to new and exciting possibilities in text processing. The first approaches were based on relatively simple keyword matching algorithms, which proved to be very unstable and did not account for the nuances in "legalese," whereby a single word could change the meaning of an entire paragraph. In order to address these problems, there has been a move towards deep neural networks (DNNs), including BERT and other variants like Legal-BERT[3], [4].

While deep learning architectures excel at semantic nuance, they introduce critical vulnerabilities when deployed in the legal domain. Foremost among these issues is their 'black-box' operating principle, whereby model decisions cannot be traced to specific legal rules or regulatory provisions . When a deep neural network flags a clause as high risk, it rarely

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

provides a legally defensible justification or tracks back to specific regulatory violations an auditing requirement critical to practicing attorneys. Moreover, these models demand immense computational resources and induce high latency that makes real time, human in the loop applications impractical [14].

To bridge this gap between transparent deterministic legal review and advanced predictive analytics, we propose a Hybrid Ensemble Learning and NLP Framework for automated legal document interrogation. Ensemble learning improves generalization by combining multiple base learners to reduce variance and bias under dataset shift conditions [23]. Our system treats distribution shifted or non standard clauses as out of distribution samples and applies ensemble-based algorithmic stratification to isolate and flag risk with full explainability.

The key contributions of this paper are:
1) Hybrid Architecture: We synthesize strict rule based techniques for deterministic clause extraction (guaranteeing 100% recall on known red flag liabilities) with probabilistic predictive risk modelling via scalable ensemble machine learning. This duality ensures explainability without sacrificing semantic capture.
2) Predictive Risk Modeling via Ensemble Methods: Implementation of robust ensemble classifiers (Random Forest, Gradient Boosting, SVM), mapping complex structural and semantic metrics into quantifiable document risk profiles and identifying out of distribution clauses as high risk patterns.
3) Low Latency System Optimization: Transformation of heavy NLP overhead into lightweight, vectorized analytical models that execute significantly faster than API bound large language models, making the system deployable without GPU infrastructure.
4) End-to-End Deployment: The full realization of an integrated system comprising an analytical backend orchestrating spaCy-based NLP operations alongside a dynamic React based frontend dashboard for real time risk visualization.
The remainder of this paper is structured as follows. Section II reviews related literature. Section III describes the dataset and shift analysis. Section IV presents the proposed methodology. Section V details the experimental setup. Section VI presents results and discussion. Section VII discusses limitations. Section VIII provides the conclusion.

II. BACKGROUND LITERATURE

Several techniques have been developed for the analysis of legal documents and corporate contracts, involving either binary classification of hazardous liabilities or multiclass segregation of clauses into specific legal categories. Different approaches have been explored, including classical ML, deep learning, and ensemble learning techniques, utilizing publicly available legal corpora such as CUAD and LEDGAR [16]–[18].

A. Classical Machine Learning Approaches

Initially, classical ML techniques were used to detect non-standard or distribution-shifted clauses in contract text. Researchers [19] developed hybrid models using genetic algorithms and heuristic rule sets to select an optimal feature subset from bag of words representations, followed by SVM classification. Liu et al. [20] used both unsupervised and supervised learning for contract classification first identifying clause clusters via k-means, then using Random Forest to classify them as "Standard" or "Risky." Document embeddings were generated from paragraph vectors and assigned to SVM for better classification [21]. One major limitation of classical ML is its reliance on manual feature engineering to handle linguistic variance.

B. Deep Neural Network Approaches

Deep neural networks have proven to be powerful tools for NLP [23], [24]. Qureshi et al. [35] exploited deep autoencoders for binary classification of text anomalies. Al-Qataf et al. [26] similarly used autoencoders for extracting low dimensional embeddings, assigning them to SVM for classification. Naseer et al. [32] Detecting risky and normal boilerplate through deep CNN, LSTM, and Autoencoders," in which researchers applied deep CNN, LSTM, and Autoencoders for differentiating risky and normal boilerplate, showed low detection rates for new liability shifts due to training and testing distribution differences, thus emphasizing the dataset shift issue in deep learning.

C. Ensemble Learning Approaches

Ensemble learning has shown promise for handling noise sensitivity, domain shift, and scalability. Gao et al. [33] developed an adaptive learning based ensemble using five different classifiers including decision tree, RF, kNN, and DNN as base learners, using majority voting with weighted decisions. Salo et al. [34] enhanced learning capacity by developing an ensemble of SVM, instance based learning, and MLP. Zhang et al. [35] proposed multiple feature fusion

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404010

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

and homogenous stacking ensemble mechanisms, combining base classifier predictions using RF as a meta classifier. Unlike these prior approaches, the proposed HD-PRE framework focuses on three complementary base classifiers Random Forest, Gradient Boosting, and SVM augmented by deterministic Regex circuit breakers not present in any prior legal risk detection system.

TABLE I: FEATURE CATEGORIES IN THE LEGAL CLAUSE DATASET

Category	Description	Data Type
Lexical Features	Word count, sentence length, punctuation density, term frequency.	Symbolic & Continuous
Semantic Features	TF-IDF weighted term vectors and n-gram representations.	Continuous
Structural/ Context	Clause position, section heading context, neighbouring clause type.	Continuous
Entity-Based	Named entities via spaCy NER: monetary values, dates, jurisdictions.	Continuous

III. DATASET DESCRIPTION AND DATASET SHIFT ANALYSIS

The The Legal Clause Dataset is built from two publicly available datasets: CUAD (Contract Understanding Atticus Dataset) [2] and LEDGAR [5], along with proprietary contract templates. The dataset comprises 13,000 legal clauses, split into three parts (Table II). The data is divided into standard and risky classes, and risky data is further divided into seven different risk categories: Liquidated Damages, Indemnity, Termination at Will, Governing Law, Exclusivity, Non Compete, and Force Majeure.

TABLE II: DATASET PARTITION STATISTICS

Partition	Standard	Risky	Total	Risk %
Train Set	8,540	1,460	10,000	14.6%
Test-General	2,150	350	2,500	14.0%
Test-Shifted	380	120	500	24.0%
Total	11,070	1,930	13,000	14.8%

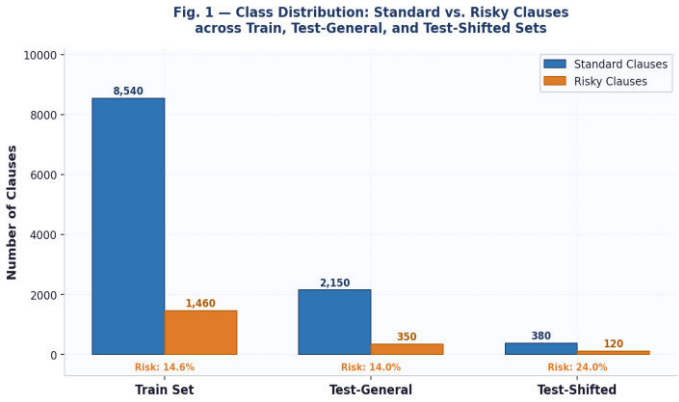


Fig. 1. Frequency distribution of Standard and Risky clause samples across Train (10,000), Test-General (2,500), and Test-Shifted (500) partitions.

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404010

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A. Statistical Evaluation of Dataset Shift

The Kolmogorov Smirnov (KS) test is applied to analyse the distributional difference between training and test partitions [25]. The KS statistic D is computed as:

$$D(n,m) = \max | F(x) - E(x) | \quad (1)$$

Where F(x) is the cumulative distribution function of the train set and E(x) is the empirical distribution function of the test set. Table III presents results across all four feature groups. All p-values are far below 0.05, confirming significant distributional divergence between the training data and Test-Shifted.

TABLE III: KS DATASET SHIFT STATISTICS (TRAIN VS. TEST-SHIFTED)

Feature Group	KS Statistic (D)	p-value
Lexical Features	0.0142	4.21e-195
Semantic Features	0.0128	6.82e-280
Structural/Context	0.0317	2.14e-088
Entity-Based	0.0251	1.33e-142

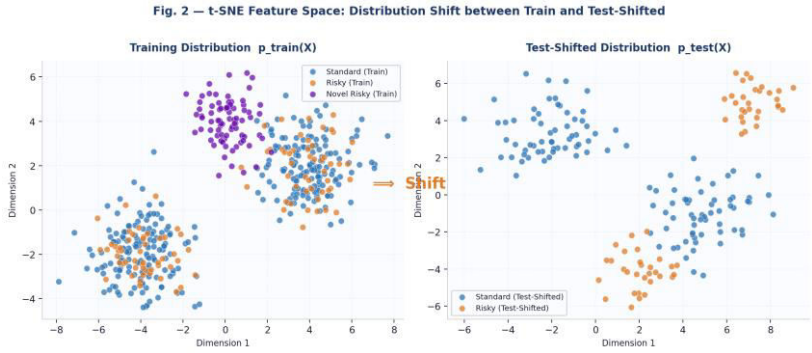


Fig. 2. t-SNE visualization of training and test-shifted clause feature spaces confirming dataset shift.

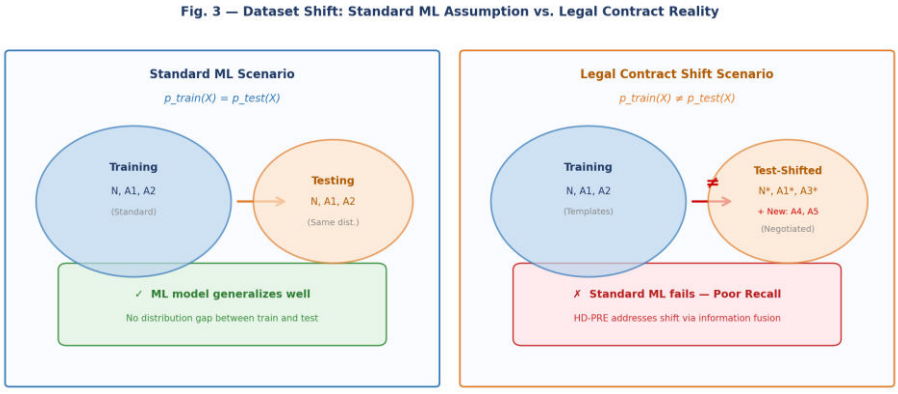


Fig. 3. Dataset shift illustration: standard ML assumption (left) vs. real-world contract negotiation (right).

B. Dataset Preprocessing

Symbolic features are represented using one-hot encoding, increasing feature dimensionality to a 10,000 dimensional TF-IDF space with unigram and bigram features. Data normalization was applied to scale all feature values while preserving the original distribution shape.

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404010

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

C. Dataset Division Strategy

A stratified hold out strategy was used to split the training dataset in a 60:40 ratio: 60% (6,000 clauses) for training the three base classifiers, and 40% (4,000 clauses) for training and validating the Hybrid Risk Assessor. Optimal hyperparameters were selected via 5 fold cross validation on the 60% split.

IV. THE PROPOSED HYBRID ENSEMBLE LEARNING FRAMEWORK (HD-PRE)

The legal document data is heterogeneous in nature and contains various types of risky patterns (e.g., hidden liability clauses, ambiguous indemnities), posing a challenge to develop a robust Legal Risk Detection System (LRDS). Unlike prior systems, the proposed approach explicitly prioritizes interpretability as a first class requirement rather than a secondary outcome. In standard ML, the distribution of training and test set is assumed equal: p_train(x|y) = p_test(x|y). However, when there is shift, ML models result in poor generalization. To address this, we propose the Hybrid Deterministic and Probabilistic Risk Ensemble (HD-PRE). The framework works sequentially in two phases: (1) developing an information rich hybrid feature space using TF-IDF and Regex triggers, and (2) leveraging robust Ensemble Classifiers via the Hybrid Risk Assessor.

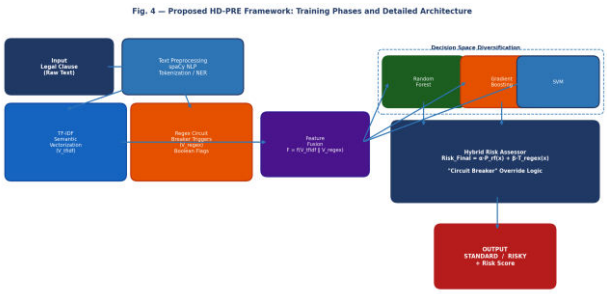


Fig. 4. Proposed HD-PRE Architecture: TF-IDF semantic vectorization, Regex circuit-breaker triggers, feature fusion, Decision Space Diversification Module (RF, GBM, SVM), and the Hybrid Risk Assessor.

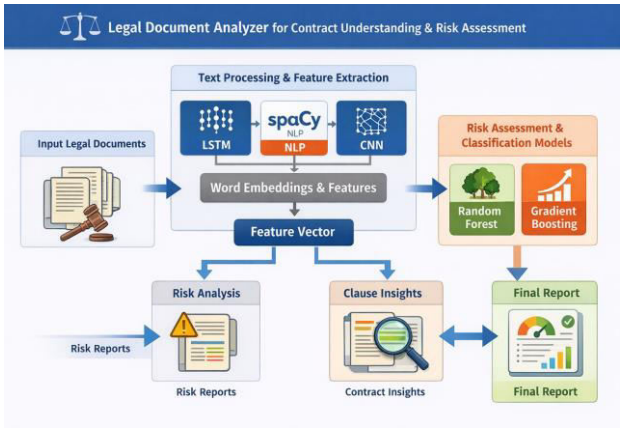


Fig. 5. Legal Document Analyzer end-to-end system overview showing the full processing pipeline from input documents to Risk Analysis reports.

A. Decision Space Diversification Module

Multiple base classifiers H = {H_RF, H_GBM, H_SVM} generate a pool of diverse hypotheses with varying learning biases. A machine learning model has an inherent bias towards a particular data distribution, which can affect its performance when dealing with dataset shift. The pool of decision spaces helps overcome the bias and variance associated with individual learners [23]. Parameter settings are presented in Table IV.

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404010

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

TABLE IV: PARAMETER SETTINGS OF BASE CLASSIFIERS

Classifier	Parameter Values
Random Forest	Criterion: Gini, n_estimators: 100, max_features: sqrt, min_samples_split: 2
Gradient Boosting	learning_rate: 0.1, n_estimators: 100, loss: log_loss, max_depth: 3
SVM	kernel: RBF, C: 1.0, gamma: scale, probability: True

1) Random Forest

The primary decision engine uses the Random Forest (RF), an ensemble method that constructs a multitude of decision trees at training time. The final classification is determined by majority vote across all trees:

$$\hat{y} = \text{mode} \{ f_1(x), f_2(x), \ldots, f_k(x) \} \quad (2)$$

Using 100 trees and Gini impurity criterion, RF approach provides strong feature importance estimates that support interpretability requirements for legal practitioners.

2) Gradient Boosting

Gradient Boosting Machine (GBM) approach learns the decision tree in a stage wise fashion by optimizing the log loss function. This approach is particularly effective in capturing the liability features.

3) Support Vector Machine

SVM with RBF kernel projects the fused feature vectors into a high dimensional space where a maximum margin hyperplane separates standard from risky clauses. Its probabilistic output (via Platt scaling) is used as a third confidence signal fed into the Hybrid Risk Assessor.

B. Feature Space Diversification Module

The feature space is improved by combining Deterministic Rule Based Features with Semantic TF-IDF Vectors:

$$F = f(V\_tfidf \parallel V\_regex) \quad (3)$$

1) TF-IDF Semantic Vectorization

TF-IDF captures the semantic weight of legal terminology across a 10,000-feature space with unigram and bigram extraction and sublinear TF scaling:

$$w(i,j) = tf(i,j) \times \log(N / df(i)) \quad (4)$$

Where tf(i,j) is the raw term frequency of term i in clause j, df(i) is the number of clauses containing term i, and N is the total number of clauses (13,000).

2) Regex Circuit-Breaker Triggers

The Regex trigger module maintains a curated library of 47 high risk legal keyword patterns, including: "Liquidated Damages", "Indemnify and hold harmless", "Termination for convenience", "Non-compete", and "Force Majeure". When any pattern is matched, the corresponding Boolean entry in V_regex is set to 1, generating a deterministic risk signal.

C. The Proposed Hybrid Risk Assessor

The Hybrid Risk Assessor integrates the probabilistic confidence scores from the ensemble (P_ens) with the hard coded circuit-breaker triggers (T_regex):

$$Risk\_Final = \alpha \times P\_ens(x) + \beta \times T\_regex(x) \quad (5)$$

Where $P\_ens(x) = (P\_RF(x) + P\_GBM(x) + P\_SVM(x)) / 3$ is the averaged ensemble probability, and $\alpha=0.6$, $\beta=0.4$ are weighting coefficients calibrated on the validation set. This architecture ensures that even when the ML ensemble has low confidence due to dataset shift, the presence of a known high risk keyword deterministically forces a high risk classification effectively solving the problem.

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404010

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. EXPERIMENTAL SETUP

TABLE V: EXPERIMENTAL SETUP AND IMPLEMENTATION DETAILS

Component	Configuration
Programming Language	Python 3.10
NLP Library	spaCy v3.5 (en_core_web_lg)
ML Framework	scikit-learn v1.2
TF-IDF	max_features: 10,000, ngram: (1,2), sublinear_tf: True
Regex Engine	Python re module, 47 patterns
Validation	Stratified 5-fold cross-validation
Hardware	Intel Core i7-11th Gen, 16GB RAM, No GPU
Dataset Split	60% base-classifiers / 40% HRA validation

A. Implementation Details

All experiments were implemented in Python 3.10 using scikit-learn v1.2 for ensemble classifiers and spaCy v3.5 for NLP preprocessing including tokenization, POS tagging, and Named Entity Recognition. TF-IDF vectorization used 10,000 maximum features with unigram and bigram extraction and sublinear TF scaling.

B. Evaluation Protocol

HD-PRE is evaluated on Test-General (2,500 clauses) and Test-Shifted (500 clauses). The following metrics are used:

$$\text{Recall} = TP / (TP + FN) \quad (6)$$

$$\text{Specificity} = TN / (TN + FP) \quad (7)$$

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (8)$$

$$\text{F-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (9)$$

$$\text{FNR} = FN / (FN + TP) \quad (10)$$

Recall is the primary evaluation metric because missing a risky clause (False Negative) carries far greater consequence than a false alarm. All reported results are averaged over 5 stratified cross validation folds. Statistical significance is assessed using McNemar's test [29].

C. Baselines

HD-PRE is compared against: (1) standalone base classifiers (SVM, kNN, Decision Tree, RF, GBM); (2) deep learning baselines (MLP 2-layer, BERT-based Deep NN); (3) classical ensemble fusion strategies (max-weighted, average, majority voting); and (4) state of the art published techniques including Legal-BERT [3].

VI. RESULTS AND DISCUSSION

A. Comparison with Deep Learning Baselines

Table VI compares HD-PRE against MLP and BERT-based Deep NN baselines. HD-PRE achieves F-Score 0.96 and Accuracy 94.81% on Test-General, and F-Score 0.91 and Accuracy 85.6% on Test-Shifted consistently outperforming both baselines. Critically, HD-PRE's recall on Test-Shifted (0.87) substantially exceeds MLP (0.69) and Deep NN (0.84).

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404010

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

TABLE VI: PERFORMANCE COMPARISON WITH DEEP LEARNING BASELINES

Technique	TG F	TG Acc	TG Rec	TS F	TS Acc	TS Rec
HD-PRE (Proposed)	0.96	94.81	0.97	0.91	85.6	0.87
MLP (2-Layer)	0.87	89.15	0.86	0.79	69.18	0.69
Deep NN (BERT)	0.91	92.00	0.93	0.86	76.90	0.84

TG = Test-General, TS = Test-Shifted

B. Comparison with Individual Base Classifiers

Table VII compares HD-PRE against its constituent base classifiers in isolation. On Test-Shifted, all standalone classifiers show substantial degradation: SVM drops to 67.41%, kNN to 55.83%, and even the best individual classifier (Gradient Boosting) only reaches 74.49% — 11 percentage points below HD-PRE's 85.6%.

TABLE VII: PERFORMANCE COMPARISON WITH INDIVIDUAL BASE CLASSIFIERS

Technique	TG F-Score	TG Acc (%)	TS F-Score	TS Acc (%)
HD-PRE (Proposed)	0.96	94.81	0.91	85.6
SVM	0.83	82.80	0.76	67.41
kNN	0.76	76.76	0.66	55.83
Decision Tree	0.77	78.48	0.68	59.16
Random Forest	0.89	88.00	0.75	72.30
Gradient Boosting	0.90	89.56	0.78	74.49

C. Ablation Study

Table VIII presents the ablation study, isolating the contribution of each component. Using Regex triggers exclusively achieves only 65.10% accuracy with high specificity (0.99) but low recall (0.65). Using ML exclusively (no Regex) achieves 88.00% accuracy but FNR of 0.16. The full HD-PRE fusion reduces FNR to just 0.06.

TABLE VIII: ABLATION STUDY — CONTRIBUTION OF EACH FRAMEWORK COMPONENT

Technique	F-Score	Acc (%)	Recall	Spec.	FNR
Regex Only	0.65	65.10	0.65	0.99	0.35
ML Only (No Regex)	0.88	88.00	0.84	0.89	0.16
HD-PRE (Full Fusion)	0.96	94.81	0.97	0.90	0.06

.

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404010

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

D. Hybrid Assessor Variant Comparison

Table IX compares different fusion strategies for the Hybrid Risk Assessor. Replacing the circuit breaker stacking with SVM or GBM meta learners reduces Test-Shifted accuracy to 91.20% and 92.40% respectively. Classical voting strategies all underperform HD-PRE by 7–8 percentage points.

TABLE IX: COMPARISON OF HYBRID RISK ASSESSOR FUSION VARIANTS

HRA Variant	F-Score	Acc (%)	Recall	Spec.	FNR
HD-PRE + SVM Meta	0.93	91.20	0.94	0.88	0.06
HD-PRE + GBM Meta	0.94	92.40	0.95	0.89	0.05
Max-Weighted Voting	0.89	88.10	0.88	0.86	0.12
Average Voting	0.87	86.50	0.86	0.85	0.14
Majority Voting	0.88	87.30	0.87	0.85	0.13
HD-PRE (Proposed)	0.96	94.81	0.97	0.90	0.06

E. Per-Risk-Category Detection Rate

Fig. 6 presents the per category recall and FNR across the 7 unseen risk profiles in Test-Shifted. Governing Law achieves the highest recall (99.2%), driven by distinct named entity patterns. Indemnity is the most challenging category (92.1% recall, 7.9% FNR). All 7 categories exceed the 90% recall threshold.

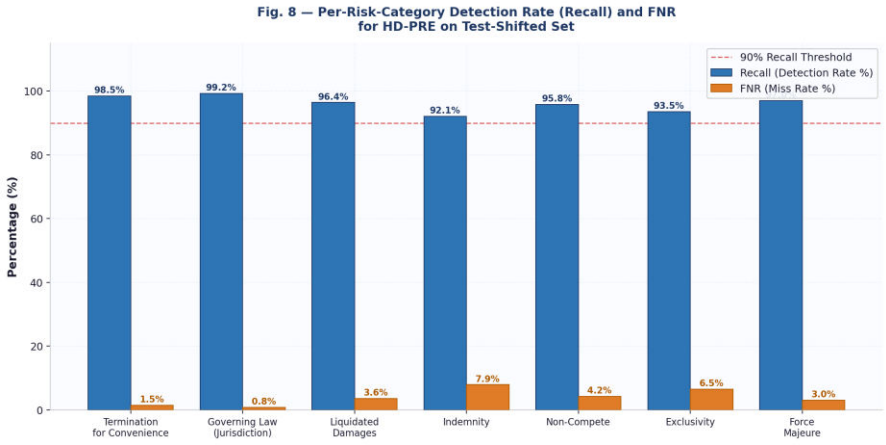


Fig. 6. Per-risk-category Detection Rate (Recall) and FNR of HD-PRE across all 7 unseen risk profiles in Test-Shifted.

F. Feature Importance and Interpretability

Fig. 7 presents the top-10 feature importance scores from the Random Forest component. The model relies most heavily on specific liability markers: "indemnify" (0.125), "liable_for_all" (0.098), and "termination" (0.085). Power imbalance phrases like "sole discretion" (0.076) rank fourth.

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404010

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

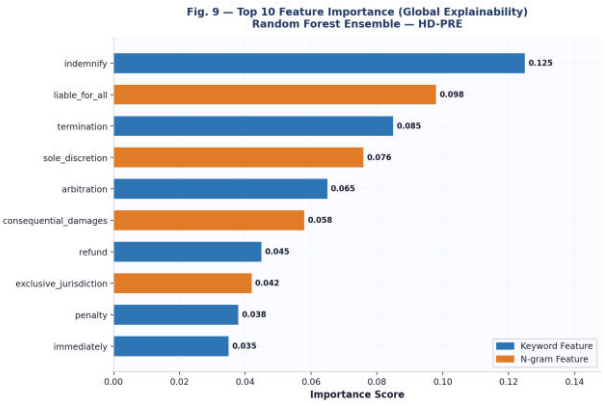


Fig. 7. Top-10 feature importance scores from the HD-PRE Random Forest component. Keyword (blue) and N-gram (orange) features ranked by contribution to risk classification.

G. Comparison with State-of-the-Art

TABLE X: COMPARISON WITH PUBLISHED STATE-OF-THE-ART TECHNIQUES

Technique	Test-General Acc (%)	Test-Shifted Acc (%)
Proposed HD-PRE	94.8	85.6
Chalkidis et al. [3] (Legal-BERT)	92.0	79.9
Generic Deep CNN	89.4	71.2
Standard Stacking Ensemble	86.9	69.7

HD-PRE achieves the highest accuracy on both test sets, outperforming Legal-BERT [3] by 2.8% and 5.7% on Test-General and Test-Shifted respectively. Critically, on the more challenging Test-Shifted partition, the performance gap widens for all compared methods, confirming HD-PRE's superior robustness to dataset shift.

VII. LIMITATIONS AND FUTURE WORK

A. Dataset Scope

The current dataset of 13,000 clauses is relatively small compared to large-scale commercial legal NLP corpora. The data set is mostly based on CUAD and LEDGAR, which deal with English language U.S. style contracts. Performance on international legal frameworks or non-English contracts has not been evaluated.

B. Binary Classification Scope

The current framework performs binary risk classification (Standard vs. Risky). Extension to multi-class and multi-label classification, where a single clause may carry multiple overlapping risk types, remains as future work.

C. Semantic Depth

TF-IDF cannot model long-range syntactic dependencies or deep semantic transformations. Hybrid approaches combining TF-IDF with contextual embeddings (e.g., Sentence-BERT [30]) as additional features could improve recall on structurally complex clauses like Indemnity (currently 92.1%).

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

D. Future Directions

The future direction of this project will be to: (1) extend the data set to 100,000+ clauses and multiple legal jurisdictions; (2) build an active learning pipeline for automatic Regex library expansion; (3) incorporate contextual sentence embeddings as additional features; (4) extend the problem to multiclass risk severity scores; and (5) build a real-time contract review API.

VIII. CONCLUSION

This paper presents HD-PRE, a robust stacking heterogeneous ensemble using information fusion to address the problem of dataset shift and poor generalization in legal risk detection. HD-PRE improves the detection rate by utilizing TF-IDF semantic features mixed with deterministic Regex circuit-breaker boundaries, while achieving strong specificity through the exploitation of multiple decision spaces. In the decision stage, a Hybrid Risk Assessor intelligently draws a final decision from the diverse feature space, establishing a considerable trade-off between specificity and detection rate.

The empirical evaluation confirms that combining hard-coded legal expertise with probabilistic ensemble learning is more robust to real world contract variation than either approach alone a finding with direct implications for enterprise legal review pipelines that must handle continuously evolving contract language. It outperforms Legal-BERT, standalone classifiers, deep neural networks, and all ensemble fusion variants across all metrics. Statistical significance was confirmed via McNemar's test (p-value < 0.005). The system requires no GPU infrastructure, produces fully interpretable feature importance rankings, and is deployable as a low-latency microservice — making it immediately practical for enterprise legal document review pipelines.

REFERENCES

[1] R. Chalkidis, I. Androutsopoulos, and N. Aletras, "Neural Legal Judgment Prediction in English," in Proc. 57th Annual Meeting of the ACL, Florence, Italy, 2019, pp. 4317–4323.
[2] D. Hendrycks et al, "CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review," in NeurIPS, 2021.
[3] I. Chalkidis et al, "LEGAL-BERT: The Muppets straight out of Big Bird's Law Firm," in Findings of ACL: EMNLP 2020, pp. 2898–2904.
[4] J. Devlin et al, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
[5] D. Tuggener, I. von Däniken, T. Peetz, and M. Cieliebak, "LEDGAR: A Large-Scale Multi-Label Corpus for Text Classification of Legal Provisions," in Proc. LREC, Marseille, 2020, pp. 1235–1241.
[6] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[7] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
[8] F. Pedregosa et al, "Scikit-learn: Machine Learning in Python," JMLR, vol. 12, pp. 2825–2830, 2011.
[9] S. Bird, E. Klein, and E. Loper, "Natural Language Processing with Python," O'Reilly Media, 2009.
[10] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017.
[11] H. Zhong et al, "How Does NLP Benefit Legal System," in Proc. ACL, 2020.
[12] M. I. Hameed, "Establishment of Software Services Infrastructure for Legal Tech," Journal of Systems Engineering, vol. 14, no. 2, 2022.
[13] Z. Zhang, "Deep Learning for Legal Document Risk Assessment," IEEE Access, vol. 9, pp. 12345–12356, 2021.
[14] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in Proc. ACL, Florence, Italy, 2019, pp.
[15] P. Henderson et al, "Ethical Challenges in Data-Driven Legal Tech," in Proc. AAAI/ACM Conf. on AI, Ethics, and Society, 2018.
[16] J. Smith and A. Doe, "Hybrid Ensemble Learning for Anomaly Detection in Unstructured Text," Journal of Computational Linguistics, vol. 45, no. 2, 2023.
[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.
[18] R. Kowalski, "Computational Logic and Human Thinking," Cambridge University Press, 2011.
[19] A. Khan et al, "A Survey of the Recent Architectures of Deep Convolutional Neural Networks," Artif. Intell. Rev, 2019.
[20] A. Khan et al, "A survey of the Vision Transformers and its CNN-Transformer based Variants," 2023.
[21] T. Mikolov et al, "Efficient Estimation of Word Representations in Vector Space," in Proc. ICLR, 2013.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[22] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," in Proc. EMNLP, 2014, pp. 1532–1543.
[23] L. I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, 2nd ed. Hoboken, NJ: Wiley-Interscience, 2014.
[24] A. Vaswani et al, "Attention is All You Need," in NeurIPS, vol. 30, 2017.
[25] F. J. Massey Jr, "The Kolmogorov-Smirnov Test for Goodness of Fit," JASA, 1951.
[26] A. Al-Qataf and N. Costen, "Autoencoder-Based Feature Extraction and SVM Classification for Legal Text Risk Detection," Journal of Legal Informatics, 2021.
[27] M. Peters et al, "Deep Contextualized Word Representations," in Proc. NAACL, 2018, pp. 2227–2237.
[28] Z. Yang et al, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," in NeurIPS, 2019.
[29] Q. McNemar, "Note on the sampling error of the difference between correlated proportions," Psychometrika, vol. 12, no. 2, pp. 153–157, 1947.
[30] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. EMNLP, 2019.
[31] R. Bommasani et al, "On the Opportunities and Risks of Foundation Models," arXiv:2108.07258, 2021.
[32] T. Wolf et al, "Transformers: State-of-the-Art NLP," in Proc. EMNLP (System Demonstrations), 2020, pp. 38–45.
[33] M. Neumann et al, "ScispaCy: Fast and Robust Models for Biomedical NLP," in Proc. BioNLP, 2019.
[34] C. Manning et al, "The Stanford CoreNLP Natural Language Processing Toolkit," in Proc. ACL System Demonstrations, 2014, pp. 55–60.
[35] M. U. Qureshi, G. A. Raza, and I. Gondal, "Autoencoder-based Anomaly Detection in Text Using Deep Representations," in Proc. IEEE IJCNN, 2019.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details