Future Technology
Open Access Journal
ISSN 2832-0379

November 2026| Volume 05 | Issue 04 | Pages 25-33

Journal homepage: https://fupubco.com/futech

https://doi.org/10.55670/fpll.futech.5.4.3

Future Publishing LLC

Article

AI-driven liability assessment framework for autonomous systems integrating machine learning with EU and Indian regulatory models

Rishi Srivastava[1]*, Sadaf Fahim[2], Garima Singh[3], Koyel Roy[4], Arunanshu Dubey[5], Rahul Tiwari[1]

[1]Atal Bihari Vajpayee School of Legal Studies, Chhatrapati Shahu Ji Maharaj University, Kanpur, Uttar Pradesh, India
[2]Chanakya National Law University, Patna, Bihar, India
[3]Faculty of Law, Jagran Lakecity University, Bhopal, Madhya Pradesh, India
[4]Symbiosis Law School, Symbiosis International (Deemed) University, Hyderabad, Telangana, India
[5]Institute of Legal Studies & Research, GLA University, Mathura, Uttar Pradesh, India

ARTICLE INFO

Article history:
Received 01 February 2026
Received in revised form
15 June 2026
Accepted 17 July 2026

Keywords:
Artificial intelligence, Liability signals, Autonomous systems, Legal text classification, Multi-label classification

*Corresponding author
Email address:
rishisrivastava@csjmu.ac.in

DOI: 10.55670/fpll.futech.5.4.3

ABSTRACT

Autonomous systems pose liability issues when automated decisions contribute to harm, failure, or non-compliance with regulations. Current legal AI studies primarily focus on classification, retrieval, question answering, and judicial prediction, and little has been done to translate legal and regulatory texts into structured signals that convey liability-relevant information. This research paper proposes a decision-support model for categorizing liability cues under European Union and Indian law. The source datasets were MultiEURLEX and IndicLegalQA, and a rule-based proxy mapping was developed to assign five labels: damage, defect, liability, regulation, and risk. This model is based on word-level and character-level TF-IDF, jurisdiction encoding, and a One-vs-Rest Linear Support Vector Machine classifier for multi-label classification. The final best configuration had an overall 91.33% subset accuracy, 96.08% Micro-F1, 87.81% Macro-F1, and 0.0201 Hamming Loss, compared to the best configurations of Logistic regression, Linear SVM, and Rand Forest. The framework aids initial legal and regulatory analysis and does not decide liability on doctrinal grounds or substitute for proficient legal analysis.

1. Introduction

Autonomous systems are increasingly used in transportation and robotics, automated decision-making, and digital governance. Their increasing autonomy raises challenging questions of liability when system decisions contribute to damage, system breakdowns, or regulatory non-compliance. These queries are particularly tricky in cross-jurisdiction practices since legal responsibilities, risk expectations, compliance requirements, and accountability frameworks vary across jurisdictions. Legal artificial intelligence has advanced through multi-label classification of legal documents, such as MultiEURLEX, which can classify complex legal documents into more than two categories [1]. The use of Indian legal datasets has similarly broadened the scope of legal AI by enabling computational analysis of jurisdiction-specific question-answer content [2]. Special-purpose language models like LEGAL-BERT also indicate that legal NLP is sensitive to legal terminology, institutional language, and meaning [3]. Even with these advancements, the current legal AI research focuses on broad legal NLP tasks. Recent surveys highlight legal document classification, legal information retrieval, question answering, summarization, and predictive analytics as the key research domains in legal NLP [4]. Judicial prediction studies are another example of machine learning's potential to uncover patterns in legal text and case decisions [5,6]. But these studies are not directly relevant to the issue of structuring liability-relevant signals for autonomous agents from legal and regulatory texts. This gap is significant because general text classification alone is insufficient to enable autonomous system governance. There is a need for techniques to detect signals in legal texts related to risk, defect, damage, liability, regulation, and accountability, useful to regulators, developers, compliance teams, and researchers. Research in legal NLP has demonstrated that legal information access can be greatly facilitated by leveraging structured computational methods at scale [7] and that argumentation-based approaches are crucial for establishing links among legal claims, duties, facts,

and consequences in legal reasoning [8]. However, little computational research exists that translates regulatory text into liability-signal categories for cross-jurisdictional, autonomous-system governance. To bridge this gap, the present study proposes an AI-based liability-signal classification framework using the European Union Legal Data and the Indian Legal Data. The framework is not a legal formulation, nor does it impinge upon judicial logic or legal analysis by experts. Rather, it maps the label to the legal and regulatory texts using a rule-based approach, extracts features from the texts using TF-IDF, adds the jurisdiction field, and classifies the texts into liability-relevant categories using multi-label machine learning models.

Transparency is at the core of this task. The outputs from liability-related tasks should not be used as black-box predictions, particularly in legal and regulatory settings. Explainable AI research aims to make it easy for users to understand the "how" and "why" behind a system's prediction, as well as the appropriate actions to take [9]. Moreover, legal text categorization cannot be weak, since legal documents might be ambiguous, overlapping, and use jurisdiction-specific terminology, all of which are essential for ensuring the reliability of regulatory analysis via machine learning [10]. We think that this research has four novel characteristics. First, it builds a cross-jurisdictional taxonomy of the liability-signals used in EU and Indian legal texts. Secondly, it translates current legal information into multi-labels with liability-related categories. Third, it compares liability-signal classification machine-learning models against classical approaches in each jurisdiction. Finally, it suggests a workflow for decision support that could enable legal signals to support autonomy-system governance. This study aims to:

- To construct a liability-signal taxonomy relevant to autonomous systems.
- To map EU and Indian legal-regulatory texts into multi-label liability categories.
- To evaluate classical machine-learning models for cross-jurisdictional liability-signal classification.
- To propose a decision-support workflow for using classified regulatory signals in autonomous-system governance.

2. Literature review

The shift in attitude in the literature on AI and autonomous systems goes beyond operational efficiency to encompass governance and accountability. In the context of cloud robotics, learning-based methods show that machine learning can assist autonomous systems' decision-making under technical constraints, but this line of research primarily focuses on network offloading and system performance rather than legal responsibility or liability modeling [11]. This situation will leave a gap between technical optimization research and the regulatory questions arising from harm caused by autonomous systems or their failure to meet legal obligations. The area of AI governance is directly tackling this regulatory vacuum. One of the key strengths of socio-legal governance research is its focus on the ethics and accountability mechanisms needed to ensure AI systems are used responsibly and in compliance with regulations and best practices. Explainability research also demonstrates that automated decisions should be explainable, especially when someone needs to explain or argue a decision [12,13]. These studies confirm the necessity of an unambiguous classification of liability signals but fail to provide a specific, operational computational pipeline for mapping legal text into liability-relevant classes. The area of legal liability scholarship that has been gaining traction recently is the allocation of responsibility, the role of human involvement, and the interplay between liability and risk management rules. The European liability literature offers examples that confirm this, and that AI liability simply cannot be distinguished from the observer's share of responsibility and other aspects of risk governance [14]. A further analysis of experts' proposals on liability demonstrates that the legal framework continues to accommodate technological autonomy and accountability [15]. Ex-post tort liability suggests that liability could provide an incentive to be safer when using AI, particularly in high-risk settings [16]. The contributions are important not only legally, but also doctrinally and with a policy focus, rather than data.

Another field of research focuses on the legal status of AI systems or on the liability of the developer, operator, manufacturer, or user [17]. The emergence of autonomous vehicle technology has also created a conflict between the new technology and the current liability regimes [18]. Further, comparative literature on the law of autonomous vehicles holds that laws must also adjust to situations that autonomous systems can generate with a mix of software, sensors, and automated decision-making [19]. While these studies are important in that they delineate the legal issue of autonomous-system liability, they have yet to operationalize liability analysis by machine-learning classification of regulatory text. The European Union (EU) is increasingly highlighting the need for structured AI liability governance through regulatory efforts. The proposed AI Liability Directive seeks to address issues of civil liability arising from AI-related harm and evidentiary challenges [20]. However, such regulatory tools remain as “rules of the road.” They do not offer (single) computational approaches to identifying signals for liability across large texts of law. In the field of machine learning, there is literature that enables classification systems to be made more interpretable. This is particularly relevant in high-stakes legal contexts, where model outputs must be subject to review and should not be interpreted as model decisions [21]. This is consistent with the current study's choice to model transparent features and evaluate multi-labels. In general, literature presents 3 gaps. Legal liability is rarely modeled in technical AI studies, and scalable classification pipelines are not typically linked to legal AI liability studies or explainable AI research to cross-jurisdictional regulatory liability analysis. The contribution of this study includes the use of the liability-signal taxonomy, EU and Indian legal datasets, the TF-IDF approach for feature extraction, jurisdiction encoding, Linear SVM classification, and a decision-support workflow for autonomous-system governance to address these gaps.

3. Methodology

3.1 Research design

In this work, a computational legal text classification approach is employed to identify signals of liability in legal materials from the EU and India. The framework is not meant

to prescribe legal liability or to be a substitute for judicial or expert legal reasoning. Rather, it categorizes legal and regulatory text into structured liability-signal classes to aid initial review by regulators, compliance teams, legal researchers, and autonomous-system developers. One unit of analysis is a legal/regulatory text instance. Multiple "signal" issues can relate to a single text, and the task is designed as a multi-label classification problem. The result is a predicted set of labels (damage, defect, liability, regulation, risk), along with a liability-relevance level, to facilitate interpretation and decision support. The workflow is:

Input legal/regulatory text → preprocessing → liability-label mapping → feature extraction → jurisdiction encoding → model training → multi-label prediction → liability-signal scoring → human interpretation

The pipeline is represented as:

$$D = \{(x_i, j_i, y_i)\}_{i=1}^{n} \quad (1)$$

$$x_i \rightarrow P(x_i) \rightarrow F(x_i) \oplus J(j_i) \rightarrow \hat{y}_i \rightarrow S_i \quad (2)$$

where $D$is the dataset, $x_i$is a legal text instance, $j_i$is the jurisdiction, $y_i$is the true multi-label vector, $P(x_i)$ is preprocessing, $F(x_i)$ is TF-IDF feature extraction, $J(j_i)$ is jurisdiction encoding, $\oplus$denotes feature concatenation, $\hat{y}_i$ is the predicted label vector, and S_i is the liability-relevance score.

3.2 Data collection and label construction

Two public legal datasets were chosen: MultiEURLEX (EU legal documents) and IndicLegalQA (legal question-answer records from India) [22, 23]. The datasets were chosen to include both inter-jurisdictional legal text sources. But both datasets were designed specifically to assess liability for an autonomous system. This study thus considers them as legal-text sources and builds the liability-signal labels by a rule-based proxy-mapping procedure. Table 1 summarizes the number of records in the final data set (21,301). This comprised EU records from MultiEURLEX (18842) and Indian records from IndicLegalQA (2459). These data were then coded to five liability-signal labels: damage, defect, liability, regulation, and risk.

Liability label construction and mapping: Liability labeling was not done by the expert by hand, but was done using keywords and the respective rules. All records were first mapped to a common schema with text, jurisdiction, dataset source, and split information. Second, five types of liability were identified based on various legal concepts related to autonomous-system governance: damage, defect, liability, regulation, and risk. Third, for each category, a set of keywords was prepared. Fourth, one or more labels were added to each text instance as the relevant keyword group was found in the text. Fifth, records with no liability indicator were omitted. It is important to realize that this mapping is a proxy mapping for locating signals of liability in text, as listed in Table 2. It does not create "legal liability." The labels were also created by the rules, so inter-annotator agreement, such as Cohen's kappa, could not be used. Future validation needs include expert annotation and agreement testing.

3.3 Sampling and data split

The training set, validation set, and test set for the mapped corpus were created with a fixed random number seed of 42, using a 70% / 15% / 15% split. A split maintained the multi-label nature of the data and the representation from both jurisdictions. Finally, there were 14,910 records for training, 3,195 for validation, and 3,196 for testing.

$$D = D_{train} \cup D_{val} \cup D_{test} \quad (3)$$

where D_{train}, D_{val}, and D_{test}denote the training, validation, and held-out test sets.

3.4 Feature extraction and model configurations

Text was represented using Term Frequency–Inverse Document Frequency (TF-IDF). Word-level TF-IDF used n-grams from 1 to 2, maximum features of 80,000, min_df = 2, and max_df = 0.95. Character-level TF-IDF used the char_wb analyzer with n-grams from 3 to 5, maximum features of 50,000, and min_df = 2. Jurisdiction was represented using one-hot encoding with two dimensions. The final improved feature vector therefore had 130,002 dimensions.

$$X_i = [TFIDF_{word}(x_i) \oplus TFIDF_{char}(x_i) \oplus J(j_i)] \quad (4)$$

Table 1. Dataset summary

Dataset	Jurisdiction	Samples used	Labels mapped	Train	Validation	Test
IndicLegalQA	India	2,459	5	1,721	369	369
MultiEURLEX	EU	18,842	5	13,189	2,826	2,827

Table 2. Liability label taxonomy

Label	Definition	Rationale	Example indicators
damage	Text indicating harm, injury, loss, or damage.	Damage is required in most liability assessments.	damage, harm, injury, loss
defect	Text indicating product, design, software, or operational failure.	Defect is central to product/system liability.	defect, fault, error, failure
liability	Text indicating legal responsibility or accountability.	Liability attribution is central to failure analysis.	liable, liability, responsible
regulation	Text indicating legal, regulatory, or compliance duties.	Autonomous systems operate under regulatory obligations.	regulation, directive, law, compliance
risk	Text indicating danger, hazard, uncertainty, or unsafe operation.	Risk is central to AI and autonomous-system governance.	risk, danger, hazard, unsafe

The study compared baseline and proposed configurations using One-vs-Rest classifiers, as summarized in Table 3. Logistic Regression, Linear Support Vector Machine, and Random Forest served as baseline models. The proposed feature-level hybrid model combined word-level TF-IDF, character-level TF-IDF, and jurisdiction encoding with a One-vs-Rest LinearSVC classifier. Here, "hybrid" refers to feature integration, not classifier voting, stacking, or output assembling. Class imbalance was addressed using class_weight='balanced' in Logistic Regression and LinearSVC, and balanced_subsample in Random Forest. No oversampling or synthetic data generation was applied.

3.5 Evaluation strategy

Models were evaluated on the held-out test set using multi-label metrics: subset accuracy, Micro Precision, Micro Recall, Micro-F1, Macro Precision, Macro Recall, Macro-F1, and Hamming Loss. Subset accuracy measures the exact match between the full sets of predicted and true labels. Hamming Loss measures the proportion of incorrectly predicted labels across all samples and labels.

$$SubsetAccuracy = \frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i) \quad (5)$$

$$HammingLoss = \frac{1}{nL}\sum_{i=1}^{n}\sum_{j=1}^{L} I(y_{ij} \neq \hat{y}_{ij}) \quad (6)$$

where $n$is the number of samples and $L$is the number of labels.

Statistical comparisons were conducted using McNemar's test for exact-match correctness and Wilcoxon signed-rank tests for per-sample Hamming error, with significance set at $p < 0.05$.

3.6 Ethical considerations

The framework has only a supporting role in decision-making. Recognizes textual cues that indicate liability but does not assign liability. To determine legal responsibility, one must engage in contextual legal reasoning, causation analysis, foreseeability, evidence, human oversight assessment, and jurisdiction-specific interpretation. There may be structural biases in the datasets as well. EU legal texts are not as varied as legal texts in other languages, and IndicLegalQA is a question-answering corpus rather than a dedicated liability corpus. All outputs must thus be reviewed by a human legal and compliance expert before any regulatory or practical use.

4. Results

4.1 Experimental performance

Multi-label classification metrics were used to assess the models on the held-out test set. Here, exact results are reported for the baseline, proposed, and ablation configurations, as shown in Table 4. The best overall model was the Improved Proposed Hybrid SVM, comprising word-level TF-IDF, character-level TF-IDF, and jurisdiction encoding, with a One-vs-Rest LinearSVC classifier.

Table 3. Model configurations

Configuration	Features used	Classifier	Jurisdiction
TF-IDF + Logistic Regression	Word TF-IDF	One-vs-Rest LR	No
TF-IDF + Linear SVM	Word TF-IDF	One-vs-Rest LinearSVC	No
TF-IDF + Random Forest	Word TF-IDF	One-vs-Rest RF	No
TF-IDF + Jurisdiction + LR	Word TF-IDF + jurisdiction	One-vs-Rest LR	Yes
Improved Hybrid SVM	Word TF-IDF + char TF-IDF + jurisdiction	One-vs-Rest LinearSVC	Yes
No Character TF-IDF	Word TF-IDF + jurisdiction	One-vs-Rest LinearSVC	Yes
No Jurisdiction	Word TF-IDF + char TF-IDF	One-vs-Rest LinearSVC	No

Table 4. Main model performance on the held-out test set

Model	Subset Accuracy (%)	Micro Precision (%)	Micro Recall (%)	Micro-F1 (%)	Macro Precision (%)	Macro Recall (%)	Macro-F1 (%)	Hamming Loss
Improved Proposed Hybrid SVM	91.33	95.79	96.38	96.08	87.63	88.03	87.81	0.0201
Improved Hybrid without Jurisdiction	91.33	95.77	96.38	96.07	87.67	88.03	87.84	0.0202
Baseline TF-IDF + Linear SVM	90.61	94.93	96.30	95.61	84.86	87.88	86.33	0.0226
Improved Hybrid without Character TF-IDF	90.33	95.01	96.01	95.50	85.04	86.73	85.86	0.0231
Baseline TF-IDF + Random Forest	84.51	96.48	87.90	91.99	90.36	57.23	67.61	0.0391
Baseline TF-IDF + Logistic Regression	81.32	85.41	94.32	89.64	65.15	89.73	74.58	0.0557
TF-IDF + Jurisdiction + Logistic Regression	81.26	85.35	94.29	89.60	65.07	89.67	74.50	0.0559

It achieved 91.33% subset accuracy, 96.08% Micro-F1, 87.81% Macro-F1, and 0.0201 Hamming Loss. The model without jurisdiction encoding achieved results nearly identical to the previous model, with an accuracy of 91.33%, a micro F1 of 96.07%, and a Hamming Loss of 0.0202. This means that a wealth of jurisdictional cues had already been infused into the text through vocabulary, statutory language, and document form. Bootstrap confidence intervals were also calculated for the best model. The 95% confidence interval was 0.9042–0.9230 for subset accuracy, 0.9565–0.9649 for Micro-F1, and 0.8617–0.8926 for Macro-F1.

4.2 Baseline comparisons

As observed in Figure 1 and Figure 2, the baseline comparison shows that Linear SVM has produced better performance when compared to Logistic Regression and Random Forest classifiers. The baseline model, TF-IDF + Linear SVM (90.61% subset accuracy and 95.61% Micro-F1) outperforms Logistic Regression (81.32% subset accuracy, 89.64% Micro-F1). Random Forest had a higher Micro Precision (96.48%), but the Micro Recall (87.90%) and subset accuracy (84.51%) were lower, which suggests poor exact multi-label prediction. The improvement of the proposed model over the baseline Linear SVM is modest but measurable: subset accuracy increased from 90.61% to 91.33%, Micro-F1 increased from 95.61% to 96.08%, and Hamming Loss decreased from 0.0226 to 0.0201.

4.3 Ablation study

The ablation study examined the effect of removing jurisdiction encoding and character-level TF-IDF features. Table 5 reports the exact ablation results. As shown in Figure 3, when using character-level TF-IDF, the model performance is better. The character-level features were removed, and the subset accuracy dropped to 90.33% while Hamming Loss went up to 0.0231 from 0.0201.

In contrast, the exclusion of explicit jurisdiction encoding had very little effect. One possible explanation is that word-level and character-level TF-IDF had already identified the specific words and legal structures, institutional terms, and document structure related to jurisdiction. This needs to be seen as a finding about the dataset, not as a statement that jurisdiction features are not needed in general.

4.4 Jurisdiction-wise analysis

The analysis of test records has considered them separately for the EU and India. The results are not directly comparable to the pooled result in Table 4. The average accuracy of the subsets in the held-out data set is 91.33% and is pooled. The performance on the same test set is instead reported in Table 6 when the set is split by jurisdiction. Lower values for jurisdictions may result from differences in label distributions, dataset structures, and text formats.

As shown in Figure 4, the EU subset achieved 82.00% subset accuracy and 89.84% Micro-F1, while the Indian subset achieved 75.61% subset accuracy and 86.87% Micro-F1. This performance difference may reflect the more standardized structure of MultiEURLEX compared with the question-answer format of IndicLegalQA, rather than a substantive difference in legal complexity between jurisdictions.

4.5 Liability assessment distribution

The liability assessment layer converts predicted liability signals into low-, moderate-, and high-liability-relevance categories, as summarized in Table 7. These categories are intended only as operational screening outputs for human review and should not be interpreted as legally calibrated risk thresholds or final liability findings.

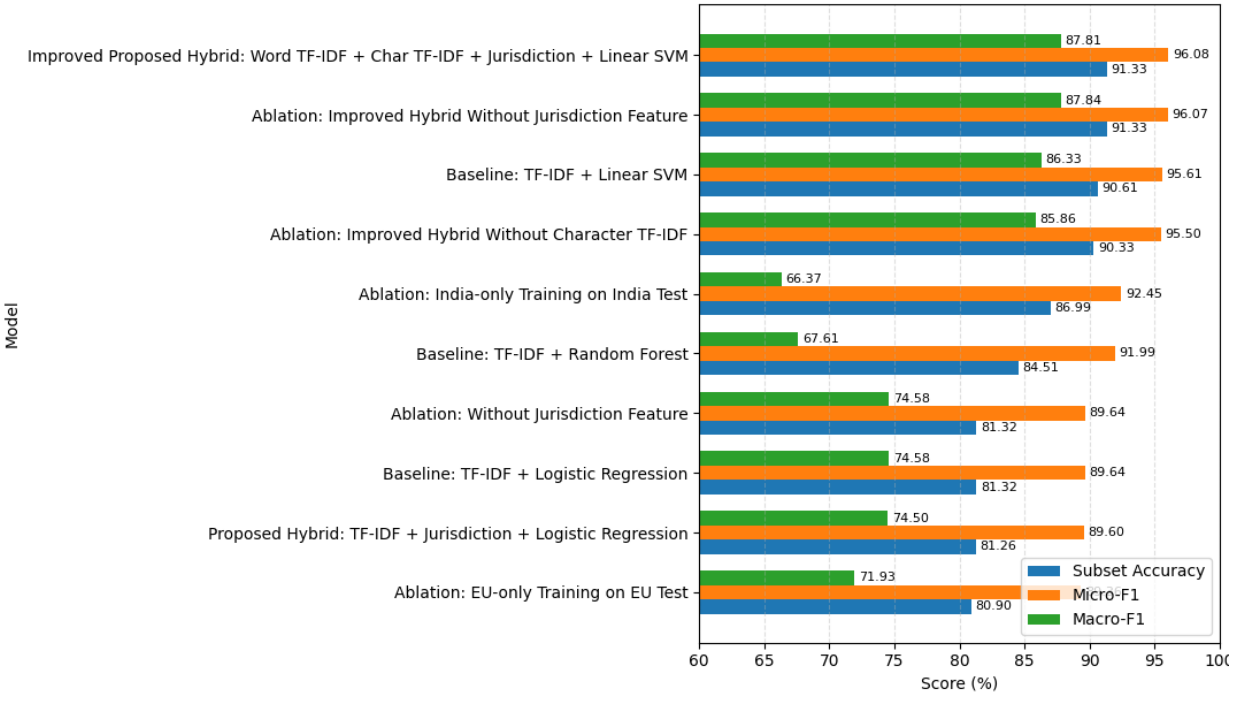

Figure 1. Model performance comparison across subset accuracy, Micro-F1, and Macro-F1

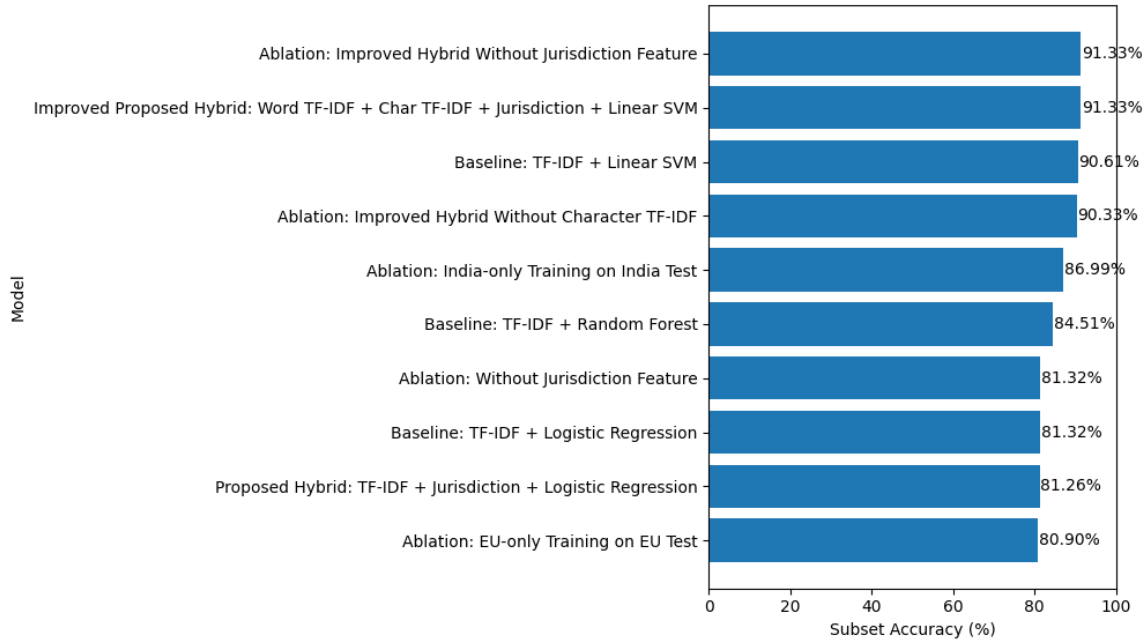

Figure 2. Subset accuracy comparison across model configurations.

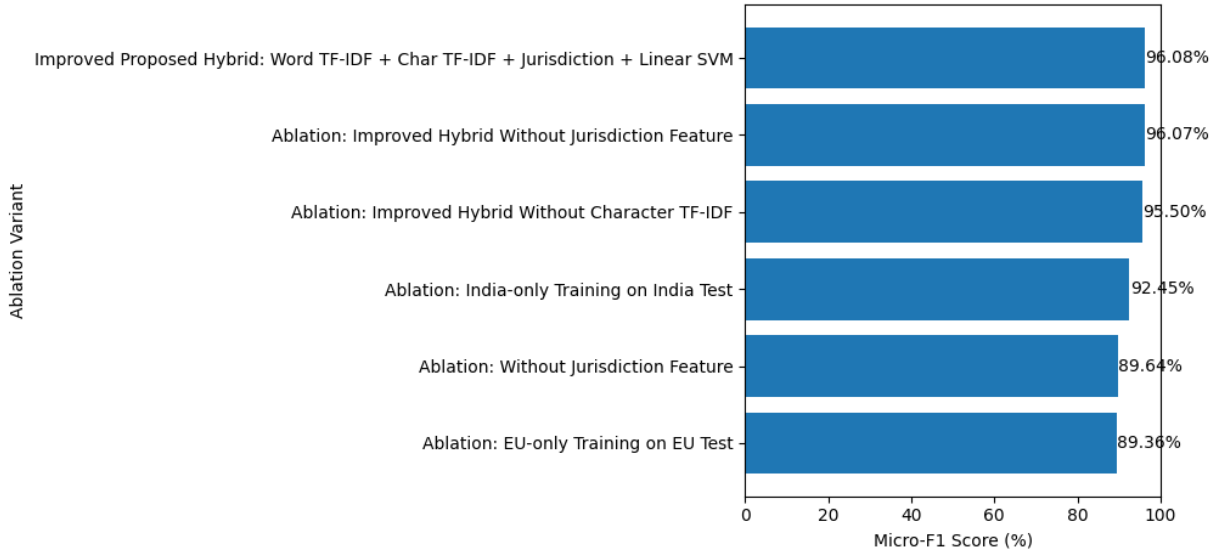

Figure 3. Ablation study results showing the contribution of model components

Table 5. Ablation study results

Variant	Removed / restricted component	Subset Accuracy (%)	Micro-F1 (%)	Macro-F1 (%)	Hamming Loss	Interpretation
Full Improved Hybrid SVM	None	91.33	96.08	87.81	0.0201	Best overall configuration
Improved Hybrid without Jurisdiction	Jurisdiction one-hot encoding	91.33	96.07	87.84	0.0202	Explicit jurisdiction encoding adds limited additional gain
Improved Hybrid without Character TF-IDF	Character-level TF-IDF	90.33	95.50	85.86	0.0231	Character features improve legal-term and subword representation
India-only Training on India Test	EU records removed	86.99	92.45	66.37	0.0314	Tests India-specific model behavior
EU-only Training on EU Test	India records removed	80.90	89.36	71.93	0.0593	Tests EU-specific model behavior

Table 6. Jurisdiction-wise performance of the proposed model

Model	Subset Accuracy (%)	Micro Precision (%)	Micro Recall (%)	Micro-F1 (%)	Macro Precision (%)	Macro Recall (%)	Macro-F1 (%)	Hamming Loss
Proposed Hybrid – EU	82.00	84.34	96.10	89.84	62.94	89.96	73.00	0.0568
Proposed Hybrid – India	75.61	99.34	77.18	86.87	78.13	68.28	72.68	0.0493

Table 7. Liability assessment distribution by jurisdiction

Jurisdiction	Low (%)	Moderate (%)	High (%)
EU	30.03	33.18	36.79
India	55.83	31.44	12.74

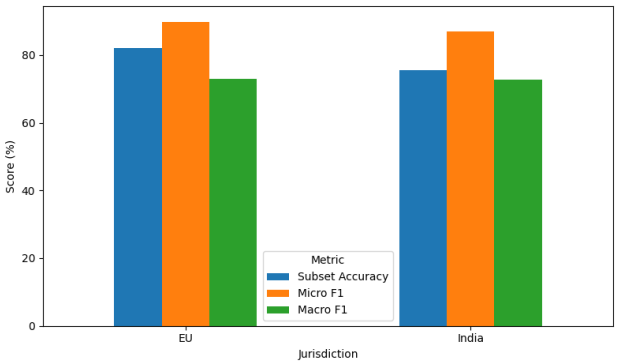

Figure 4. Jurisdiction-wise performance of the proposed model

As illustrated in Figure 5, EU records were distributed more evenly across the three liability-relevance categories, with 36.79% classified as high relevance. In contrast, Indian records were concentrated in the low-relevance category, with 55.83% classified as low and only 12.74% classified as high. This difference may reflect dataset composition and document structure rather than substantive legal differences between jurisdictions.

4.6 Statistical significance results

McNemar's test was used for exact-match correctness, and the Wilcoxon signed-rank test was used for per-sample Hamming error. The significance threshold was $p<0.05$. As shown in Table 8, the proposed model significantly outperformed Logistic Regression, Random Forest, baseline Linear SVM, and the no-character-TF-IDF ablation. However, the difference between the fully improved model and the version without jurisdiction encoding was not statistically significant. This confirms that character-level text representation contributed more to performance than explicit jurisdiction encoding in this experiment.

5. Discussion

The results suggest that classical machine-learning models can recognize signals in legal and regulatory documents relevant to liability when the task is framed as multi-label classification.

The best results were achieved by the Improved Hybrid SVM with word-based TF-IDF, character-based TF-IDF, and jurisdiction encoding. This indicates that well-designed sparse textual features can be effective for structured legal classification tasks, particularly when the target language is a limited set of liability-signal labels rather than open-ended legal reasoning. The excellent performance of SVM is consistent with the structure of the feature space. TF-IDF results in high-dimensional sparse vectors, and Linear SVM is well-suited for separating classes in high-dimensional sparse vector spaces. This is why SVM was better than Logistic Regression and Random Forest. Random Forest was less effective due to the limitations of tree-based models with sparse legal-text features, whereas Logistic Regression was competitive but less accurate at predicting multi-labels with high precision. The ablation results explain why feature design is important. Character-level TF-IDF also improved performance, since legal texts include technical terms, statutory expressions, abbreviations, and morphology that are not fully captured by word-level features. However, removing explicit jurisdiction encoding did not significantly impact the improved model. One possible explanation is that the features at the word and character levels have already indirectly reflected jurisdiction through legal vocabulary, legal institutional references, legal language patterns, and document styles. This is dataset-dependent and should not be interpreted as evidence that jurisdiction features are irrelevant to legal AI. Jurisdiction-wise performance was better for EU data than for Indian data. This change is probably related to the structure of the datasets. MultiEURLEX contains more standardized EU legal documents, whereas IndicLegalQA is a question-answer dataset that may not reflect the structure of statutory and regulatory liability provisions.

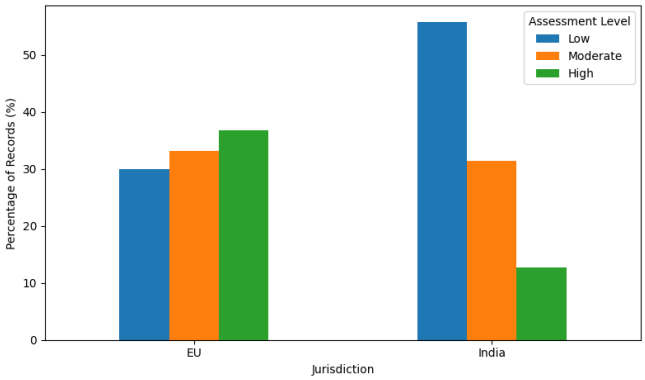

Figure 5. Liability assessment level distribution by jurisdiction

Table 8. Statistical significance tests comparing the best model against alternatives

Compared model	McNemar p-value	Wilcoxon p-value	Significant
Baseline TF-IDF + Logistic Regression	2.576208e-67	3.258245e-68	Yes
Baseline TF-IDF + Linear SVM	0.02652511	0.001109118	Yes
Baseline TF-IDF + Random Forest	6.569423e-35	1.407044e-37	Yes
TF-IDF + Jurisdiction + Logistic Regression	7.798109e-68	6.215103e-69	Yes
Improved Hybrid without Character TF-IDF	0.00160058	5.403282e-05	Yes
Improved Hybrid without Jurisdiction	1.00000000	0.6547208	No

Indian legal texts can also be more interpretive and context-dependent, making them difficult to classify with keyword-derived labels and TF-IDF features. The results are in line with and more focused than the existing legal NLP and AI-liability literature. Previous legal NLP research indicates that machine learning can be used to categorize legal documents and facilitate large-scale access to legal information [1,3,7]. But most of these works focus on general legal classification and representation rather than liability-signal modeling for autonomous systems. Likewise, AI governance and liability scholarship has focused on accountability, human oversight, and risk-based responsibility [12,14,20], yet it has typically remained doctrinal or policy-oriented. The present study ties these strands by operationalizing liability concepts as multi-label text-classification outputs.

When it comes to using the framework, it should be viewed as a screening tool and decision support system. For autonomous vehicle compliance review, a regulator or developer could input a new EU or Indian regulatory text. The system would perform pre-processing on the text, classify liability-relevant labels (e.g., risk, defect, damage, regulation, liability), assign a low, moderate, or high liability-relevance level, and mark for human inspection. The path forward is regulatory text, liability-signal classifier, label output, relevance score, human legal/compliance review, and autonomous-system governance action. The output can be used to pinpoint areas that may need further legal review, compare regulatory trends across jurisdictions, and prioritize compliance reviews. It is not meant to be a determination of legal responsibility or a substitute for legal adjudication.

Several limitations remain. First, the model is based solely on legal and regulatory text – it does not incorporate sensor logs, simulation outputs, incident reports, accident records, or operational data from autonomous systems. Second, MultiEURLEX and IndicLegalQA were not developed for autonomous-system liability assessment, and their labels are rule-based proxies rather than expert-validated liability annotations. Thirdly, there are validity issues with question-answer records as they do not necessarily correspond to liability provisions. Fourth, causation, foreseeability, negligence, proof of product defect, failure of human oversight, standards of evidence, and legal context are not evaluated. Fifth, results are based only on EU and Indian data. Lastly, TF-IDF and SVM might not capture more intricate semantic connections that transformer-based legal models can capture. We plan to further research expert annotation, extended jurisdictions, actual autonomous-system incident data, and explainable transformer-based models.

6. Conclusion

This study aimed to address the challenge of identifying signals of liability in legal and regulatory documents related to autonomous systems. It suggested an AI classification model developed on EU and Indian legal datasets, comprising rule-based liability-label mapping, word-level TF-IDF, character-level TF-IDF, Jurisdiction encoding, and multi-label machine learning. The Improved Proposed Hybrid SVM model achieved the best performance, with 91.33% subset accuracy, 96.08% Micro-F1, 87.81% Macro-F1, and a Hamming Loss of 0.0201 on the held-out test set. The ablation study revealed that character-level TF-IDF was helpful, and explicit jurisdiction encoding was helpful to an extent once richer text features were added. A significant contribution is the transparent, multi-jurisdictional classification of liability-related signals in legal texts from the EU and India. It can help regulators, compliance teams, researchers, and autonomous-system developers pinpoint legal materials that need a closer look. But the framework is only applicable to text-based legal information and does not include liability annotations from experts; it applies rule-based proxy annotations. It does not assess autonomous-system behavior, causation, foreseeability, human oversight, or evidentiary standards. Thus, it should be used only as a tool for decision-making, not as a replacement for legal adjudication or expert interpretation. Going forward, datasets labeled by experts should be used, along with more jurisdictions, legal models based on transformers, explainable deep learning, and real-life information from autonomous systems such as incident reports, simulations, and operational logs.

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically with regard to authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with policies on research ethics. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere. All survey participants provided informed consent prior to participation, and the anonymity and confidentiality of all respondents were strictly maintained throughout the research process.

Data availability statement

The manuscript contains all the data. However, additional data will be provided by the corresponding author upon reasonable request.

Conflict of interest

The authors declare no potential conflict of interest.

References

[1] I. Chalkidis, M. Fergadiotis, and I. Androutsopoulos, "MultiEURLEX: A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 6974–6996, doi: 10.18653/v1/2021.emnlp-main.559.

[2] K. Veningston and A. Mishra, "Dataset for legal question answering system in the Indian judiciary context," Data in Brief, vol. 60, Art. no. 111647, 2025, doi: 10.1016/j.dib.2025.111647.

[3] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 2898–2904, doi: 10.18653/v1/2020.findings-emnlp.261.

[4] F. Ariai, J. Mackenzie, and G. Demartini, "Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges," ACM Computing Surveys, vol. 58, no. 6, pp. 1–37, 2025. Available: https://arxiv.org/abs/2410.21306.

[5] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos, "Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective," PeerJ Computer Science, vol. 2, Art. no. e93, 2016, doi: 10.7717/peerj-cs.93.

[6] M. Medvedeva, M. Vols, and M. Wieling, "Using machine learning to predict decisions of the European Court of Human Rights," Artificial Intelligence and Law, vol. 28, no. 2, pp. 237–266, 2020, doi: 10.1007/s10506-019-09255-y.

[7] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, "How does NLP benefit the legal system: A summary of legal artificial intelligence," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5218–5230, doi: 10.18653/v1/2020.acl-main.466.

[8] T. Bench-Capon, K. Atkinson, and A. Wyner, "Using argumentation to structure e-participation in policy making," in Transactions on Large-Scale Data- and Knowledge-Centered Systems XVIII, Berlin, Germany: Springer, 2015, pp. 1–29.

[9] D. Gunning, E. Vorm, J. Y. Wang, and M. Turek, "DARPA's explainable AI (XAI) program: A retrospective," Applied AI Letters, vol. 2, no. 4, Art. no. e61, 2021, doi: 10.1002/ail2.61.

[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in Proceedings of the International Conference on Learning Representations, 2018. Available: https://arxiv.org/abs/1706.06083.

[11] S. Chinchali, A. Sharma, J. Harrison, A. Elhafsi, D. Kang, E. Pergament, E. Cidon, S. Katti, and M. Pavone, "Network offloading policies for cloud robotics: A learning-based approach," Autonomous Robots, vol. 45, no. 7, pp. 997–1012, 2021. Available: https://arxiv.org/abs/1902.05703.

[12] A. Theodorou and V. Dignum, "Towards ethical and socio-legal governance in AI," Nature Machine Intelligence, vol. 2, no. 1, pp. 10–12, 2020, doi: 10.1038/s42256-019-0136-y.

[13] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," Harvard Journal of Law & Technology, vol. 31, no. 2, pp. 841–887, 2017. Available: https://arxiv.org/abs/1711.00399.

[14] B. Botero Arcila, "AI liability in Europe: How does it complement risk regulation and deal with the problem of human oversight?," Computer Law & Security Review, vol. 54, Art. no. 106012, 2024, doi: 10.1016/j.clsr.2024.106012.

[15] A. Bertolini and F. Episcopo, "The expert group's report on liability for artificial intelligence and other emerging digital technologies: A critical assessment," European Journal of Risk Regulation, vol. 12, no. 3, pp. 644–659, 2021, doi: 10.1017/err.2021.30.

[16] G. Noto La Diega and L. C. T. Bezerra, "Can there be responsible AI without AI liability? Incentivizing generative AI safety through ex-post tort liability under the EU AI liability directive," International Journal of Law and Information Technology, vol. 32, no. 1, 2024.

[17] S. Chesterman, "Artificial intelligence and the limits of legal personality," International & Comparative Law Quarterly, vol. 69, no. 4, pp. 819–844, 2020, doi: 10.1017/S0020589320000366.

[18] G. E. Marchant and R. A. Lindor, "The coming collision between autonomous vehicles and the liability system," Santa Clara Law Review, vol. 52, no. 4, pp. 1321–1340, 2012. Available: https://digitalcommons.law.scu.edu/lawreview/vol52/iss4/6.

[19] M. L. Kubica, "Autonomous vehicles and liability law," The American Journal of Comparative Law, vol. 70, suppl. 1, pp. i39–i69, 2022.

[20] European Commission, "Proposal for a directive on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)," COM(2022) 496 final, Brussels, Belgium, 2022. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0496.

[21] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608, 2017. Available: https://arxiv.org/abs/1702.08608.

[22] I. Chalkidis, M. Fergadiotis, and I. Androutsopoulos, "MultiEURLEX dataset," Hugging Face Datasets, 2021. Available: https://huggingface.co/datasets/nlpaueb/multi_eurlex.

[23] K. Veningston and A. Mishra, "Indic legal question answering dataset," Mendeley Data, V2, 2025. Available: https://data.mendeley.com/datasets/gf8n8cnmvc/2.