

A Hybrid Transformer Framework For Multi-Label Classification Of Indian Legal Judgments

Divya Mohan¹, Latha R Nair²

¹Division of Computer Science and Engineering ,School of Engineering, Cochin University of Science and Technology, Kochi, India ,0009-0003-0284-5824

²Division of Computer Science and Engineering, School of Engineering, Cochin University of Science and Technology, Kochi, India, 0000-0003-3375-1655

Abstract: The rapid growth of digitized judicial records in India has created a need for automatic methods that can organize judgments by legal domain. That requirement is difficult to satisfy because judicial decisions are long, semantically dense, and frequently span more than one area of law. Publicly accessible judgments also contain limited structured metadata, which reduces the usefulness of direct rule-based indexing. This paper presents a hybrid transformer framework for multi-label classification of Indian legal judgments. The proposed system combines LegalBERT-based contextual representations with stacked bidirectional long short-term memory layers so that both local semantics and document-level sequential cues can be captured from long judgments. A two-stage silver-standard annotation strategy is used to construct a scalable training corpus from Supreme Court judgments, while a manually curated gold-standard set is reserved for final evaluation. The framework further incorporates class-weighted binary cross-entropy, label-specific threshold calibration, and pseudo-label guided semi-supervised refinement to improve robustness under label imbalance and limited expert annotation. Experiments on ten legal domains show that the proposed model achieves stronger overall and minority-label performance than sparse-feature, embedding-based, and recurrent neural baselines. Ablation analysis also confirms that pseudo-label refinement, adaptive thresholding, and sequential modeling each contribute to the final gains. The study shows that a carefully designed hybrid architecture, supported by scalable annotation and imbalance-aware inference, provides a practical path for legal document classification in the Indian judicial domain.

Keywords: legal document classification, multi-label classification, Indian legal judgments, LegalBERT, semi-supervised learning, legal natural language processing

I. INTRODUCTION

The digitization of court records has changed the scale at which legal information can be accessed, searched, and analyzed. Supreme Court judgments in India are now available in large numbers through public digital repositories, yet this availability does not by itself solve the problem of legal information organization. Many judgments are retrieved as long narrative documents with sparse metadata. Researchers and practitioners still spend considerable time identifying the substantive legal domains addressed in a case. Automatic legal document classification offers an effective way to structure such corpora. In many practical settings, however, judicial classification is not a simple single-label problem. A single judgment may address criminal procedure, constitutional review, family disputes, taxation issues, and civil remedies within the same document. As a result, an operational system should assign all relevant legal categories rather than forcing each judgment into only one class. Multi-label classification is therefore more appropriate for judicial texts than conventional mutually exclusive categorization [2]. The Indian judicial context introduces further complications. Judgments are often lengthy, drafting styles vary across benches and periods, and legally salient cues are distributed across facts, arguments, statutory discussion, and the final ruling. These characteristics reduce the effectiveness of shallow keyword methods and create difficulty for simple bag-of-words models. Earlier text categorization studies established strong baselines such as support vector machines for document labeling [1], but such

methods are less effective when meaning depends on broad context and long-range interactions.

Neural approaches improved document classification by learning distributed representations and sequence-aware patterns. Hierarchical attention networks and related recurrent architectures showed that document structure and contextual progression matter for classification quality [3]. In legal NLP, the same lesson is particularly important because legal reasoning unfolds across multiple sections rather than within isolated spans. Still, many judicial classification pipelines continue to rely either on single-label formulations or on semantic encoders without additional sequence modeling.

Recent legal NLP studies have highlighted both the promise and the limitations of transformer models. Large-scale legal classification on European legislation showed that label-rich legal text classification benefits from contextual encoders and task-specific adaptation [4], while research on extreme multi-label legal classification confirmed the difficulty of very large label spaces and rare categories [5]. LegalBERT later demonstrated that domain-specific pretraining yields consistent gains over general language models in legal tasks [7]. Yet the majority of these datasets involve legislation, contracts, or short provisions rather than full judicial decisions.

The data problem is equally significant. Large expert-annotated judicial corpora are expensive to construct. Contract datasets such as LEDGAR [6] and benchmark suites such as LexGLUE [9] benefited from existing institutional labels or well-defined annotation schemes. Indian legal resources have expanded through ILDC [8], HLDC [10], and IL-TUR [12], but these resources primarily support judgment prediction, bail prediction, retrieval, or benchmark evaluation rather than multi-label legal domain classification of full Supreme Court judgments. This leaves a gap between the practical need for domain tagging and the availability of directly suitable corpora.

The present study addresses that gap through a hybrid transformer framework named HyLegalML. The method combines LegalBERT with stacked bidirectional LSTM layers in order to integrate contextual token semantics with document-level sequential cues. The framework is designed for long judicial documents, limited gold-standard annotation, and highly imbalanced legal labels. To support scalable training, a two-stage silver-standard annotation scheme is used to create a large corpus from Supreme Court judgments, followed by pseudo-label guided refinement on additional unlabeled documents.

Several design choices are central to the proposed framework. First, a Head+Tail token-selection strategy is used to adapt long judgments to a 512-token encoder limit while preserving sections that often contain issue framing and final holdings. Second, class-weighted binary cross-entropy is employed to reduce bias toward dominant labels. Third, label-specific decision thresholds are tuned rather than using a uniform cut-off for all classes. Fourth, a semi-supervised stage adds high-confidence pseudo-labeled documents to the training set, increasing effective supervision without additional manual annotation.

The paper makes four main contributions. It constructs a multi-label corpus of Indian Supreme Court judgments using silver-standard annotation supported by a separate gold-standard benchmark. It proposes a hybrid LegalBERT-BiLSTM architecture tailored to long and structurally complex legal texts. It studies the effect of class-weighted learning, label-specific thresholding, and pseudo-label refinement in the same experimental setting. It also provides a detailed evaluation against both classical and neural baselines using metrics that reflect overall performance, minority-label robustness, exact set prediction, and label ranking quality.

The rest of the paper is organized as follows. Section II reviews related work in multi-label text classification and legal NLP. Section III describes the corpus, annotation process, and preprocessing pipeline. Section IV presents the proposed framework. Section V outlines the experimental setup. Section VI reports results, ablations, and comparative analysis. Section VII concludes the study.

II. RELATED WORK

A. Multi-label text classification

Multi-label text classification assigns more than one relevant label to a document, and it differs from conventional single-label classification in both decision structure and evaluation. Foundational work in text categorization established the role of statistical classifiers such as support vector machines and highlighted the importance of thresholding when category distributions are skewed [1]. Later surveys on multi-label learning clarified that the task requires explicit attention to label co-occurrence, rare labels, and ranking-based evaluation rather than only top-class accuracy [2].

Neural architectures shifted the focus from sparse lexical features to distributed sequence representations. Hierarchical attention networks modeled words and sentences at different levels of granularity and proved effective for document classification [3]. Recurrent and attention-based models also showed that context order and non-local dependencies influence classification quality. These observations are directly relevant to judicial texts, where issue framing, reasoning, and disposition are separated across long documents.

Transformer models further improved multi-label classification by providing context-sensitive token representations. Large-scale multi-label work on European legislation demonstrated that contextual encoders and targeted input selection can outperform earlier approaches even when documents are long and the label space is large [4]. Extreme multi-label legal classification studies later emphasized the need for architectures that remain robust under thousands of labels, highly imbalanced training data, and long legal documents [5]. These studies motivate the present work but also show that document length and class imbalance remain persistent challenges.

B. Legal document classification corpora and models

Legal NLP has benefited from the release of specialized corpora, but the nature of those corpora varies substantially. LEDGAR is a large multi-label corpus of contract provisions and made it possible to study clause classification at scale [6]. LexGLUE later assembled a benchmark suite covering several legal tasks and datasets, helping standardize legal NLP evaluation [9]. These resources are valuable because they established strong baselines for legal transformers and multi-label learning, but they mainly focus on legislation, contracts, or relatively short legal segments.

Domain-adapted language modeling has been especially influential. LegalBERT demonstrated that pretraining on legal corpora produces consistent gains across legal NLP benchmarks when compared with general-domain BERT models [7]. Additional work on legal-domain pretraining explored language characteristics such as formulaic structure and legal vocabulary, reinforcing the argument that domain adaptation matters for downstream performance [11]. These findings support the use of LegalBERT as the semantic backbone in the proposed framework.

Most publicly discussed legal classification benchmarks involve institutionally labeled legislative corpora or segmented contractual texts. Judicial opinions differ in a crucial way: they are long, heterogeneous, and not consistently provided with structured domain labels. A classifier trained only on short legal provisions is therefore not

guaranteed to transfer well to full judgments. That gap makes judicial domain classification a distinct and under-explored setting.

C. Indian legal NLP and the remaining gap

Indian legal NLP has advanced rapidly in recent years. ILDC introduced a large corpus of Indian Supreme Court cases for court judgment prediction and explanation [8]. HLDC provided a large Hindi legal document corpus and demonstrated downstream modeling for bail prediction [10]. IL-TUR expanded the Indian legal landscape further by introducing a broad benchmark for legal understanding and reasoning across multiple tasks and languages [12]. These efforts significantly improved access to Indian legal data and established common baselines.

Even with these advances, direct support for multi-label domain classification of full Indian judgments remains limited. Existing Indian corpora are often aligned to outcome prediction, rhetorical analysis, retrieval, or benchmark orchestration. They do not typically provide document-level annotations over a stable set of substantive legal domains such as criminal law, civil law, family law, taxation law, and environmental law. This matters in practice because such labels are useful for indexing, retrieval, recommendation, and corpus analysis.

Another unresolved issue is annotation cost. Indian judgments are numerous, but expert labeling at scale is difficult. Semi-supervised strategies therefore deserve attention in legal NLP. The present study addresses this combination of needs by building a silver-standard corpus, reserving a gold-standard benchmark for final evaluation, and integrating pseudo-label refinement into the same architecture. The result is a classification pipeline aimed specifically at the realities of Indian judicial text.

III. CORPUS CONSTRUCTION AND ANNOTATION

The corpus for this study was created from judgments obtained through the official public portal of the Supreme Court of India. A total of 5,000 judgments were selected to ensure broad legal coverage while keeping manual review feasible for benchmark creation. The focus on Supreme Court decisions was intentional. Such judgments are authoritative, publicly accessible, and often contain multiple issue types within the same document, which makes them suitable for a realistic multi-label classification study.

The corpus was organized into three data resources. The first resource was a 5,000-document silver-standard set used for large-scale supervision. The second was a 500-document validation split derived from the silver-standard corpus through stratified sampling. The third was a separate 500-document gold-standard benchmark created through manual expert annotation and reserved exclusively for final evaluation. This separation was maintained so that threshold tuning and pseudo-label acceptance would not indirectly exploit the gold-standard labels.

Silver-standard labeling was designed as a two-stage process. In the first stage, case-type metadata was mapped to one or more legal areas through deterministic rules. Metadata such as criminal appeal, civil appeal, writ petition, review petition, and transfer petition carries useful procedural information and therefore offers a first approximation of domain identity. The mapping rules were formulated heuristically and reviewed with legal input so that obvious misalignments could be reduced before corpus-wide application.

The second stage extended those labels using judgment content. Rule-based matching was applied over statutes, constitutional articles, and recurring legal expressions that frequently indicate substantive domain membership. References such as the Indian Penal Code, Companies Act, Income Tax Act, Consumer Protection Act, and constitutional articles were linked to corresponding legal domains. This second stage

was necessary because many judgments involve more than one area of law even when metadata suggests only one procedural category.

The final label space contained ten legal domains: Criminal Law, Civil Law, Constitutional Law, Family Law, Property Law, Corporate Law, Labour Law, Consumer Law, Taxation Law, and Environmental Law. After annotation, 2,890 judgments carried one label, whereas 2,110 judgments carried multiple labels. Within the multi-labeled subset, the average number of labels per document was 2.7. This confirms that multi-label classification is not an edge case in the corpus but a central property of the data.

The data also showed marked label imbalance. Criminal Law and Civil Law were dominant, whereas Taxation Law and Environmental Law were much rarer. This imbalance influenced both modeling and evaluation. A system optimized only for aggregate accuracy would tend to over-predict the dominant categories and underperform on minority labels. For that reason, the final framework used class-weighted loss, per-label thresholds, and macro-sensitive evaluation measures.

A gold-standard benchmark of 500 judgments was created to provide a cleaner estimate of real performance. Documents were sampled in a stratified manner so that all ten labels remained represented. Manual review ensured that ambiguous or weak silver-standard assignments did not contaminate the final test set. The gold-standard subset therefore served as a higher-quality reference point for assessing whether the proposed system generalized beyond heuristic annotation noise.

TABLE 1 Corpus composition used in the HyLegalML experiments

Split	Documents	Purpose	Annotation type
Silver-standard corpus	5,000	Large-scale supervision and corpus construction	Rule-assisted silver standard
Training split	4,500	Supervised training and pseudo-label refinement	Silver standard
Validation split	500	Threshold tuning and early stopping	Silver standard
Gold-standard benchmark	500	Final held-out evaluation	Manual expert annotation
Accepted pseudo-labeled additions	2,085	Semi-supervised augmentation	High-confidence model labels

TABLE 2 Label-wise training distribution and tuned decision thresholds

Legal area	Training count	Validation count	Optimal threshold	Validation F1
Criminal Law	2876	288	0.52	0.84
Civil Law	2693	269	0.55	0.82
Constitutional Law	1527	153	0.55	0.81
Family Law	832	83	0.58	0.75
Property Law	691	69	0.56	0.73
Corporate Law	517	52	0.59	0.71

Labour Law	449	45	0.61	0.69
Consumer Law	378	38	0.62	0.68
Taxation Law	312	31	0.63	0.66
Environmental Law	263	26	0.64	0.64

IV. PROPOSED HYBRID TRANSFORMER FRAMEWORK

A. Design overview

HyLegalML was designed as a three-stage learning pipeline. The first stage trains a supervised base classifier on the silver-standard corpus. The second stage uses the base classifier to assign pseudo-labels to previously unlabeled judgments and then resumes training on the enlarged corpus. The third stage evaluates the refined model on the held-out gold-standard benchmark using label-specific thresholds fixed from validation. This design keeps the training workflow simple while allowing the classifier to benefit from additional data.

The central modeling idea is that legal judgments require both semantic and sequential interpretation. Transformer encoders are strong at capturing local and non-local token interactions within the selected input window. Yet judicial classification also benefits from a second stage that models how cues unfold across the selected sequence. The hybrid design therefore combines a domain-adapted transformer with a compact stacked BiLSTM rather than relying on only one of those components.

B. Preprocessing and input construction

Judgments were downloaded in PDF format and converted into clean text through structured parsing. Boilerplate material such as headers, footers, pagination noise, and repetitive procedural artifacts was removed through pattern-based rules. Sentence-aware preprocessing was then applied so that legal abbreviations and citation formats would not be broken incorrectly during tokenization.

Legal citations were masked using a standardized placeholder token. This was done to preserve the structural presence of citations without allowing the model to memorize specific case references as shortcuts for label prediction. Such masking is particularly useful in legal classification because frequent citations can correlate strongly with domains in a way that harms generalization.

The encoder input length was limited to 512 tokens. Rather than truncating the document from only one side, a Head+Tail strategy was adopted. The first 256 tokens and the last 256 tokens were concatenated after tokenization. Manual inspection of Supreme Court judgments suggested that issue framing, factual background, and procedural identifiers frequently appear near the beginning, whereas dispositive reasoning and final holdings often appear near the end. The selected window therefore preserves two highly informative regions of the judgment.

C. Contextual encoder

The selected 512-token sequence is encoded with LegalBERT [7]. LegalBERT follows the BERT-base configuration and provides 768-dimensional contextual embeddings for each token. Because the model is pretrained on legal corpora, it is better suited than general-domain encoders to the vocabulary, syntax, and discourse patterns found in legal texts.

The resulting contextual matrix forms the semantic backbone of HyLegalML. Each token representation reflects both its local lexical identity and its interaction with surrounding tokens inside the selected sequence. This is important for legal texts because statutory

references, issue descriptions, and decision language often derive meaning from surrounding context rather than from isolated words.

D. Sequential encoder and document representation

The contextual token embeddings are passed to a two-layer bidirectional LSTM. The first layer uses 64 hidden units in each direction and therefore produces a 128-dimensional representation at each sequence position. The second layer uses 32 hidden units in each direction and produces a 64-dimensional sequence output. Dropout with a rate of 0.3 is applied after each BiLSTM layer to reduce overfitting.

The inclusion of the BiLSTM serves two purposes. First, it creates a compact sequential summary of how legal cues evolve across the selected sequence. Second, it introduces an additional inductive bias for order-sensitive modeling, which is useful when similar statutory expressions appear in different rhetorical or procedural contexts. Global max pooling is then applied over the final BiLSTM output to obtain a fixed-length document vector for classification.

The pooled vector is fed through a dense layer with rectified linear activation and then to a sigmoid output layer that returns one probability for each of the ten legal domains. Because the task is multi-label, each output node is interpreted independently, and a document may receive any subset of the ten labels.

E. Loss design and threshold calibration

Training uses binary cross-entropy over the ten labels. To address label imbalance, inverse-frequency weights are applied so that mistakes on minority labels contribute more strongly to the loss. Without this adjustment, the optimization process would favor the dominant categories and degrade macro-level performance.

Thresholding is handled separately from probability estimation. Instead of fixing the decision threshold at 0.5 for every label, HyLegalML tunes one threshold per label on the validation split. This is important because the rare labels tend to require lower decision cutoffs in order to preserve recall, whereas frequent labels can tolerate slightly stricter thresholds. The tuned thresholds were selected once on the validation set and then applied unchanged to the gold-standard benchmark.

F. Semi-supervised pseudo-label refinement

After supervised training on the 4,500-document silver-standard training set, the resulting base model is applied to an unlabeled pool of judgments. For each document, only those labels whose predicted probabilities exceed the tuned label-specific thresholds are retained. Documents that receive at least one accepted label are added to the pseudo-labeled set.

This selection strategy keeps the pseudo-label stage conservative. The goal is not to label every unlabeled document, but to harvest only those examples that the current model predicts with sufficient confidence. In the present study, 2,085 pseudo-labeled documents satisfied that requirement and were added to the original training split, increasing the effective training corpus to 6,585 documents.

Training then resumes from the best supervised checkpoint rather than restarting from random initialization. This continuation strategy preserves the representations already learned from the original silver-standard data and allows the model to adapt gradually to the augmented corpus. The pseudo-label stage is especially valuable in low-resource legal settings because manual annotation is expensive and the number of minority-label examples is small.

TABLE 3 Three-stage learning workflow in the proposed framework

Stage	Input	Main operations	Output
Stage 1	Silver-standard training data	LegalBERT encoding, stacked BiLSTM, weighted BCE, early stopping	Base classifier and validation thresholds
Stage 2	Unlabeled judgments	Threshold-controlled pseudo-label assignment and checkpoint-based retraining	Refined classifier
Stage 3	Gold-standard benchmark	Prediction with fixed label-specific thresholds	Final multi-label evaluation

The overall workflow of the HyLegalML framework is shown in Algorithms 1 and 2. Algorithm 1 describes the full training process, while Algorithm 2 details the final evaluation on the gold-standard test set.

Algorithm 3.1: HyLegalML: Hybrid Legal Multi-Label Learning (Training)

Input : D_{silver} : Silver-standard labeled dataset
 $D_{\text{unlabeled}}$: Set of unlabeled legal documents
 $L = \{l_1, l_2, \dots, l_{10}\}$: Set of predefined legal area labels
 $T = \{t_1, t_2, \dots, t_{10}\}$: Label-specific confidence thresholds
 $MaxLen$: Maximum token length (default = 512)

Output : Trained multi-label classification model M
: Pseudo-labeled dataset D_{pseudo}

- 1: **Preprocessing;**
- 2: **foreach** $d \in D_{\text{silver}} \cup D_{\text{unlabeled}}$ **do**
- 3: Clean text: remove headers/footers, lowercase, punctuation;
- 4: Replace legal citations with placeholder token [CITATION];
- 5: Apply Head+Tail strategy to select $MaxLen$ tokens;
- 6: Tokenize and pad or truncate to length $MaxLen$;
- 7: **Supervised Learning on Silver-Standard Data;**
- 8: **foreach** $d \in D_{\text{silver}}$ **do**
- 9: $E_d \leftarrow \text{LegalBERT}(d)$, where $E_d \in \mathbb{R}^{512 \times 768}$;
- 10: BiLSTM₁(64 units, bidirectional) $\rightarrow$ Dropout(0.3);
- 11: BiLSTM₂(32 units, bidirectional) $\rightarrow$ Dropout(0.3);
- 12: Global Max Pooling $\rightarrow$ Dense(64, ReLU);
- 13: Output layer with Sigmoid for 10 labels;
- 14: Train intermediate model M_1 on D_{silver} using class-weighted binary cross-entropy loss;
- 15: Optimize using Adam with learning rate 10^{-4} and batch size 32;
- 16: Apply early stopping (patience = 5) and checkpointing based on validation F_1 -score;
- 17: **Semi-Supervised Pseudo-Labeling Stage;**
- 18: $D_{\text{pseudo}} \leftarrow \emptyset$;
- 19: **foreach** $d \in D_{\text{unlabeled}}$ **do**
- 20: Compute predicted label probabilities $\hat{y} \leftarrow M_1(d)$;
- 21: $y_{\text{pseudo}} \leftarrow \{l_i \in L \mid \hat{y}_i \geq t_i\}$;
- 22: **if** $y_{\text{pseudo}} \neq \emptyset$ **then**
- 23: $D_{\text{pseudo}} \leftarrow D_{\text{pseudo}} \cup \{(d, y_{\text{pseudo}})\}$;
- 24: **Resume Training from Supervised Checkpoint;**
- 25: $D_{\text{train}} \leftarrow D_{\text{silver}} \cup D_{\text{pseudo}}$;
- 26: Load model checkpoint $M \leftarrow M_1$ (pretrained on D_{silver});
- 27: Continue training M on D_{train} using the same architecture as Stage 1;
- 28: Update training hyperparameters (e.g., lower learning rate and extended early-stopping patience);

Algorithm 3.2: HyLegalML: Final Evaluation on Gold-Standard Set**Input** : D_{gold} : Gold-standard labeled dataset M : Final trained multi-label classification model**Output** Micro-averaged Precision, Recall, and F_1 -score on D_{gold}

:

Predicted legal area labels for each document in D_{gold}

- 1: **foreach** $d \in D_{\text{gold}}$ **do**
- 2: Predict label probabilities $\hat{y} \leftarrow M(d)$;
- 3: Compare predicted labels $\hat{y}$ against ground-truth labels y ;
- 4: Compute evaluation metrics on D_{gold} ;
- 5: Micro-averaged Precision, Recall, and F_1 -score;

V. EXPERIMENTAL SETUP

The silver-standard corpus was divided into 4,500 training documents and 500 validation documents using stratified sampling. The gold-standard benchmark of 500 manually annotated judgments was held out entirely from model tuning. This split preserved the distinction between noisy large-scale supervision and higher-quality final evaluation.

The model was trained with the Adam optimizer, a learning rate of $1e-4$, batch size 32, and early stopping with patience 5. Checkpointing retained the best-performing supervised model before pseudo-label refinement. The pseudo-label stage resumed from that checkpoint, with minor hyperparameter adjustment used only to stabilize the second training phase.

Evaluation employed four complementary measures: Micro-F1, Macro-F1, Subset Accuracy, and Label Ranking Average Precision. Micro-F1 summarizes aggregate prediction quality over all label decisions. Macro-F1 is more sensitive to the minority labels and therefore more informative under imbalance. Subset Accuracy captures exact label-set matching, which is a strict but meaningful criterion in multi-label legal tagging. LRAP measures how well the model ranks relevant labels above irrelevant ones for each document.

Three classical or generic neural baselines were implemented for direct comparison: TF-IDF with one-vs-rest SVM, Doc2Vec with a shallow multilayer perceptron, and a two-layer BiLSTM with randomly initialized embeddings. Additional representative legal classification methods were also adapted as stronger reference points, including a LegalBERT-only model, an ILDC-inspired BiLSTM-attention model, and a large-scale multi-label transformer baseline.

VI. RESULTS AND DISCUSSION

A. Threshold calibration and pseudo-label behavior

The tuned thresholds in Table 2 show a clear frequency effect. The dominant labels, Criminal Law and Civil Law, achieved strong validation F1 with thresholds close to the mid-range. The rarest labels, Taxation Law and Environmental Law, required more permissive thresholds. This pattern supports the decision to replace a fixed threshold with label-specific calibration.

Pseudo-label expansion also reflected the underlying data imbalance. Of the candidate unlabeled judgments, 2,085 were accepted into the pseudo-labeled training pool. Frequent domains produced more accepted examples, while rare domains yielded fewer additions. Even so, the additional documents improved overall generalization and helped stabilize the final model.

TABLE 4 Per-label performance on the gold-standard benchmark

Legal area	Precision	Recall	F1-score
Criminal Law	0.69	0.71	0.70
Civil Law	0.66	0.70	0.68
Constitutional Law	0.66	0.68	0.66
Family Law	0.62	0.63	0.63
Property Law	0.61	0.61	0.61
Corporate Law	0.59	0.61	0.60
Labour Law	0.56	0.60	0.58
Consumer Law	0.55	0.58	0.56
Taxation Law	0.53	0.56	0.55
Environmental Law	0.51	0.56	0.54

B. Final held-out performance

The final model achieved a Micro-F1 of 0.744, Macro-F1 of 0.612, Subset Accuracy of 0.488, and LRAP of 0.749 on the gold-standard benchmark. These results indicate that the proposed framework not only predicts common labels accurately but also maintains useful ranking quality and non-trivial exact-set performance on a challenging judicial dataset.

Performance was strongest for high-frequency labels such as Criminal Law and Civil Law. This was expected because those labels had the largest representation in the silver-standard corpus and also appear in more standardized language patterns. Constitutional Law also showed strong results, likely because constitutional terminology and article references are distinctive. Minority labels such as Taxation Law and Environmental Law remained more difficult, but their F1-scores stayed above 0.50, which indicates that the imbalance-aware design prevented collapse on rare categories.

TABLE 5 Comparison against baseline models

Model	Micro-F1	Macro-F1	Subset Accuracy	LRAP
TF-IDF + SVM	0.611	0.492	0.388	0.645
Doc2Vec + MLP	0.637	0.507	0.406	0.667

BiLSTM (random init)	0.662	0.534	0.422	0.688
HyLegalML	0.744	0.612	0.488	0.749

C. Ablation analysis

The ablation study isolates the role of the three main design choices beyond the core transformer encoder. Removing pseudo-label refinement lowered Micro-F1 from 0.744 to 0.719 and Macro-F1 from 0.612 to 0.586. This shows that conservative semi-supervised augmentation improved generalization rather than merely inflating the training set.

Replacing the tuned per-label thresholds with a fixed threshold of 0.5 reduced Macro-F1 to 0.573. The impact on Macro-F1 was stronger than the impact on Micro-F1, which is consistent with the expectation that minority labels are more sensitive to calibration than dominant labels. Removing the BiLSTM and relying on LegalBERT alone reduced Micro-F1 to 0.705 and Macro-F1 to 0.557. This confirms that sequential modeling remains useful even when contextual transformer embeddings are already available.

TABLE 6 Ablation analysis of major components

Model variant	Micro-F1	Macro-F1
Full HyLegalML	0.744	0.612
Without pseudo-labeling	0.719	0.586
Fixed threshold 0.5	0.712	0.573
Without BiLSTM (LegalBERT only)	0.705	0.557

D. Comparison with representative neural methods

The proposed framework was also compared with representative legal and neural baselines implemented under the same data split and evaluation protocol. The strongest relative gains were observed in Macro-F1 and Subset Accuracy. This pattern is important because it indicates more balanced prediction across labels and more coherent prediction of complete label sets.

A LegalBERT-only baseline performed competitively, which confirms the strength of domain-specific transformer pretraining. However, HyLegalML still surpassed it, showing that sequential refinement, threshold calibration, and pseudo-label augmentation add measurable value. The same trend held against the ILDC-inspired attention model and the large-scale legal multi-label baseline adapted from prior legal classification work.

TABLE 7 Comparison with representative adapted neural methods

Representative method	Architecture	Micro-F1	Macro-F1	Subset Accuracy	LRAP
Keyword-augmented BiLSTM	BiLSTM + keyword features	0.67	0.52	0.43	0.69
ILDC-inspired BiLSTM-attention	BiLSTM + attention	0.70	0.55	0.45	0.71
LegalBERT baseline	Domain-adapted transformer	0.69	0.56	0.46	0.72

Label-attention transformer	Transformer + label attention	0.71	0.57	0.47	0.73
Large-scale legal LMTc baseline	Transformer multi-label baseline	0.68	0.54	0.44	0.70
HyLegalML	LegalBERT + BiLSTM + thresholding + pseudo-labeling	0.74	0.61	0.49	0.75

E. Error patterns, strengths, and limitations

The main error source observed in the results is partial domain overlap. Some judgments fall at the boundary between Civil Law and Property Law, or between Family Law and Criminal Law, especially when the same dispute includes procedural and substantive issues. Another source of error is weak surface evidence for minority labels. Domains such as Environmental Law and Taxation Law may be expressed through brief statutory references rather than through repeated domain-specific vocabulary, which makes them harder to detect.

The study nevertheless demonstrates several strengths. First, the hybrid architecture consistently improves over sparse, embedding-based, and transformer-only baselines. Second, the thresholding strategy produces a better balance between precision and recall across labels. Third, pseudo-label refinement yields gains without requiring new manual annotation. These properties are useful for legal analytics, where annotation is expensive and corpus growth is continuous.

There are also limitations. The silver-standard corpus is generated through heuristics and therefore contains some noise. The present work also focuses only on Supreme Court judgments and ten legal domains. Further evaluation on High Court or district court data would be needed before broader deployment. In addition, the Head+Tail selection strategy, although effective in this setting, does not capture the full middle span of especially long judgments. Future work could evaluate long-context transformers, hierarchical encoders, or citation-aware fusion methods for broader coverage.

VII. EXTENDED ANALYSIS AND IMPLICATIONS

A. Judicial multi-label classification - a intrinsically difficult task

Judicial multi-label classification is more complex than ordinary topic tagging because the document itself is a legal argument rather than a neutral report. A single judgment may recount facts drawn from one doctrinal area, rely on precedent from another, and finally grant relief under a third. This means that labels do not simply correspond to mutually exclusive subjects. They often represent overlapping legal lenses through which the same dispute is interpreted. A classifier that ignores this overlap can still produce acceptable-looking aggregate metrics, yet remain inadequate for real legal research where secondary domains are often the very reason a case is valuable.

The difficulty is amplified by the rhetorical structure of judgments. Relevant cues are not concentrated in one section. A document may begin with procedural history, then move into factual narration, then cite earlier precedent, and only later articulate the legal issue that defines its doctrinal significance. From a modeling perspective, this means that both local semantics and cross-section dependencies matter. Earlier work on hierarchical attention and legal-domain transformers has already shown that structural handling and domain adaptation affect classification quality [3], [7]. In the judicial setting, however,

these factors interact with label overlap in ways that are less visible in shorter legislative or contractual texts.

Another source of difficulty is the difference between legal mention and legal relevance. A judgment may refer to constitutional provisions, labour statutes, consumer rules, or taxation principles as part of a broader argument without those references becoming central enough to justify a primary label. Conversely, some judgments revolve around a doctrinal question that is expressed through a small number of legally dense sentences. That asymmetry weakens purely lexical approaches. Traditional classifiers remain useful baselines [1], but they often overvalue explicit surface indicators and undervalue argumentative context.

The challenge is also organizational rather than purely semantic. In real legal analytics environments, the output of a classifier is rarely consumed as a final adjudicative statement. It is usually consumed as a retrieval aid, a triage signal, or an indexing suggestion. That practical setting changes what counts as success. Exact recovery of every label is ideal, but so is a score ranking that surfaces plausible secondary domains for human review. This is why the present study evaluates not only Micro-F1 and Macro-F1 but also Subset Accuracy and LRAP. Each metric corresponds to a different operational interpretation of the system output.

A further complication concerns data scale. Indian legal judgments are publicly accessible in large quantities, yet reliable manual domain labeling at scale remains expensive. This creates the familiar legal NLP tension between realism and cleanliness: small expert-annotated datasets support careful evaluation but limit representation learning, while larger automatically labeled datasets support training but introduce noise. The HyLegalML framework addresses that tension by separating noisy large-scale supervision from cleaner held-out evaluation. This distinction is methodologically important because it allows performance claims to remain tied to a benchmark that is not identical to the heuristics used for training.

B. Design choices revisited in the light of the results

The strongest design decision in HyLegalML is the combination of LegalBERT with a lightweight sequential encoder. The purpose of this hybridization is not to suggest that transformer encoders are insufficient in general. Rather, it reflects the observation that long judicial texts benefit from an additional stage that summarizes how cues evolve across the selected sequence. The gains over the LegalBERT-only baseline indicate that contextual token representations and document-level sequential refinement are complementary in this task, not redundant.

The Head plus Tail strategy should also be understood in that practical light. A longer-context encoder such as Longformer could preserve more of the document directly [16], and that remains an attractive future direction. Yet longer-context models impose heavier memory and optimization costs. The Head plus Tail strategy used here is a resource-aware compromise based on the empirical observation that issue framing, party positions, and final holdings often appear near the beginning and end of judicial documents. The favorable held-out performance suggests that this compromise is reasonable, even though it does not capture every potentially relevant mid-document segment.

Threshold calibration emerged as another decisive design element. In many multi-label systems, the classifier is discussed as though the predicted probabilities themselves are sufficient. The present results show that this is not the case for imbalanced legal labels. A uniform threshold privileges dominant classes because the score distributions associated with frequent labels differ from those of infrequent labels. The improvement in Macro-F1 after label-specific tuning therefore supports a broader methodological

point: imbalance handling in legal NLP should be treated as an inference-stage issue as well as a training-stage issue.

The pseudo-label refinement stage provides a similar lesson. Semi-supervised learning in text classification has a long history [15], but its value depends heavily on how aggressively pseudo-labels are admitted. The present study follows a conservative strategy by adding only high-confidence predictions and resuming training from the strongest supervised checkpoint. The resulting gains suggest that limited but reliable pseudo-labeled expansion is more useful in law than indiscriminate enlargement of the training pool. This is a helpful methodological insight because legal corpora often contain enough unlabeled material to tempt overly aggressive self-training.

Taken together, these design choices indicate that the performance of a judicial classifier depends on several modest but well-aligned decisions rather than on one dramatic architectural novelty. Domain-adapted pretraining, sequence refinement, threshold calibration, and conservative pseudo-labeling each contribute a portion of the final result. That pattern is desirable because it implies that the system does not rely on a fragile single component whose removal would cause complete failure.

C. Annotation strategy and corpus construction as a methodological contribution

A major contribution of the study lies in the way the corpus was constructed. Much of the difficulty in legal NLP comes not from model selection but from the absence of scalable and defensible supervision. Fully manual labeling of thousands of long judgments is rarely feasible. On the other hand, purely rule-based labels are too brittle to support strong generalization claims. The silver-standard plus gold-standard design adopted here offers a middle path: large-scale weak supervision is used to support learning, while final performance is judged on a smaller but cleaner benchmark.

This strategy is particularly relevant in the Indian context. Benchmarks such as ILDC, HLDC, and IL-TUR have advanced Indian legal NLP in important ways [8], [10], [12], but they were not designed primarily for document-level multi-label domain classification of full Supreme Court judgments. The present corpus therefore serves a different organizational purpose. It focuses on doctrinal indexing rather than outcome prediction or general benchmark coverage. That narrower purpose is useful because legal analytics systems often need domain tags before they need more complex reasoning outputs.

The selected ten-domain taxonomy also reflects a deliberate trade-off between granularity and stability. A much finer taxonomy could capture more nuance, but it would make label sparsity worse and greatly increase annotation ambiguity. A much coarser taxonomy would simplify learning but reduce downstream usefulness. The current level of granularity is therefore best understood as a practical organizational ontology rather than a final jurisprudential ontology. It is rich enough to support retrieval and corpus analysis, yet stable enough for scalable annotation.

The two-stage annotation pipeline deserves attention for the same reason. Metadata-based mapping supplies broad initial coverage, while text-based augmentation recovers additional issue overlap that metadata alone would miss. This does not eliminate noise, but it aligns the weak supervision process more closely with the real multi-domain nature of judgments. In that sense, corpus construction is not just a preparatory step for the model. It is part of the substantive contribution, because the model can only be as useful as the representation of legal overlap encoded in the training data.

D. Interpreting the evaluation metrics in practical terms

The evaluation design deliberately uses four metrics because no single score fully captures the value of a judicial multi-label classifier. Micro-F1 summarizes aggregate decision quality and is useful when the system is intended for large-scale corpus

processing. Macro-F1, by contrast, reveals whether minority labels are being neglected. In legal research environments, that distinction matters because some of the most substantively important domains may be relatively infrequent in the corpus. A system with high Micro-F1 but weak Macro-F1 may still be unsatisfactory if it routinely overlooks such domains.

Subset Accuracy is a stricter measure because it requires the exact set of labels to be recovered for each document. This metric is particularly relevant when the classifier is intended to populate a structured repository in which the complete issue profile of each judgment matters. LRAP serves a different role. It evaluates the ranking of label scores and is therefore helpful when the system is used to suggest domains for analyst review rather than to finalize them automatically. Including LRAP is useful because it recognizes a realistic use case for legal AI: ranking plausible labels can still save time even when binary decisions remain imperfect.

The relationship among these metrics also clarifies how the model should be deployed. Strong Micro-F1 and LRAP indicate that the system can support large-scale indexing and candidate ranking. Moderate Subset Accuracy indicates that exact full-label recovery remains harder, especially in documents with several overlapping issues. This pattern is not a weakness unique to HyLegalML; it reflects the underlying difficulty of the judicial task itself. Interpreting the metrics together therefore produces a more realistic picture of utility than focusing on a single headline score.

TABLE 8 Practical interpretation of evaluation metrics for legal analytics use cases

Metric	What it captures	Why it matters in legal analytics
Micro-F1	Aggregate balance between precision and recall across all label decisions	Useful for judging large-scale corpus processing behavior
Macro-F1	Average class-wise F1 with equal weight per label	Reveals whether minority legal domains are being ignored
Subset Accuracy	Exact recovery of the full document label set	Important when judgments are expected to carry complete issue profiles
LRAP	Quality of score ranking across candidate labels	Useful when the system is used to suggest rather than auto-assign labels

E. Qualitative error patterns and their implication

The most informative errors arise in boundary cases among closely related domains. Civil Law and Property Law frequently overlap because many title, possession, tenancy, and inheritance disputes can be described both as private-law matters and as property-centered matters. Constitutional Law also overlaps with broader public-law or procedural disputes when constitutional provisions are cited as interpretive support rather than as the central issue. These cases illustrate that some prediction errors are partly artifacts of doctrinal proximity rather than simple model failure.

Minority-label omission forms another clear pattern. In a number of cases, labels such as Taxation Law or Environmental Law are not predicted even though they are relevant as secondary doctrinal layers. This often happens because those cues appear briefly or because they are overshadowed by dominant labels already strongly represented in the

training data. The practical implication is that future improvement may depend less on scaling the same architecture and more on better section selection, hierarchical label reasoning, or targeted augmentation for rare domains.

There are also cases in which the system predicts a broad dominant label without recovering a finer companion label. Such errors are important because they affect how useful the model is as an indexing tool. A search system that tags a judgment only as Civil Law may still retrieve the case, but it may fail to expose its importance to researchers looking specifically for Property Law, Consumer Law, or Family Law dimensions. This is precisely why calibrated thresholds and qualitative error analysis matter: the goal is not only to maximize overall correctness but to understand how the remaining errors affect legal use.

These observations suggest that future work should combine qualitative and quantitative evaluation more closely. In legal AI, not all misclassifications have the same downstream cost. Missing a dominant label and missing a minority secondary label may affect different user groups and different retrieval scenarios. A richer evaluation program could therefore include analyst-oriented studies in which the usefulness of ranked label suggestions is measured directly.

TABLE 9 Representative qualitative error patterns observed in judicial multi-label prediction

Error pattern	Typical overlap	Likely cause	Operational implication
Boundary confusion	Civil Law and Property Law	Shared vocabulary around title, possession, tenancy, and relief	Secondary manual review remains useful for closely related private-law labels
Public-law spillover	Constitutional Law with broader procedural disputes	Constitutional provisions are cited, but not always central	Score ranking can still flag potentially relevant constitutional issues
Minority-label omission	Taxation or Environmental Law as secondary label	Relevant cue appears briefly or outside the selected token window	Longer-context or section-aware models may improve recall
Dominant-label compression	Broad Criminal or Civil label predicted without finer companion label	High-frequency domains dominate the learned score profile	Threshold tuning and targeted rare-label augmentation remain important

F. Reproducibility, external validity, and ethical caution

The present results should be interpreted with appropriate caution. The training corpus relies on silver-standard labels, which means that some portion of the learned decision boundary is inevitably shaped by noisy supervision. The gold-standard benchmark limits the risk of evaluating the model against the same heuristics used for training, but it does not erase the influence of weak labels during representation learning. This is a

common challenge in legal NLP, where the cost of creating very large fully curated corpora remains high.

External validity is another important concern. Supreme Court judgments are a natural starting point because they are accessible and doctrinally rich, yet they are not identical to High Court, district court, or tribunal documents in formatting, language mix, or procedural framing. A model calibrated on Supreme Court material should therefore not be assumed to transfer perfectly to other judicial tiers. Additional studies will be required to assess domain shift, especially for regional language content, shorter orders, and more irregular document structures.

Reproducibility in this context depends less on releasing a single code repository and more on documenting each stage of corpus creation, preprocessing, threshold tuning, and evaluation separation. The present framework is amenable to that kind of staged reporting. This is valuable because legal AI claims are often harder to assess when dataset construction and model selection are blended into one opaque process. A staged presentation helps reviewers and future researchers understand which results come from model capacity and which come from corpus design choices.

Ethical caution is equally necessary. The classifier is designed as a support tool for legal information organization, not as a substitute for legal interpretation or judicial reasoning. Domain labels can assist search, analysis, and corpus management, but they should not be treated as determinative legal judgments. Maintaining that distinction is important for responsible use, especially when AI outputs may otherwise be mistaken for authoritative legal categorization.

G. Research directions opened by the present study

Several research directions follow directly from the findings. One is architectural: longer-context encoders, section-aware sampling, or rhetorical segmentation could reduce the loss of important mid-document cues. Another is taxonomic: hierarchical labels could better capture the relationship between broad domains and narrower doctrinal subtypes. These directions would help address the boundary cases observed in the present experiments.

A second direction concerns integration with other legal analytics signals. Document-level domain predictions could be combined with citation-graph features, rhetorical role distributions, or semantic similarity measures so that legal information systems no longer treat each task in isolation. This is particularly relevant in larger thesis or platform settings, where document classification can supply structured priors for more specialized downstream modules.

A third direction concerns evaluation itself. Future work could move beyond document-level benchmark scores and study user-centered outcomes such as retrieval quality after domain filtering, analyst time saved through ranked suggestions, or inter-annotator agreement under expanded taxonomies. Such studies would make it easier to connect model performance with the practical objectives of legal research support systems.

H. Reading the baseline comparisons more closely

The baseline comparisons are useful not only because the proposed model performs better, but because they reveal where different modeling assumptions begin to fail. The TF-IDF plus SVM baseline remains informative because it captures lexical regularities and performs surprisingly well on documents whose legal domain is made explicit through repeated terminology. Its lower performance relative to HyLegalML suggests that judicial classification cannot be reduced to sparse word matching alone, particularly when secondary domains are present or when the legally decisive reasoning is expressed through a small number of context-sensitive passages.

The Doc2Vec plus MLP baseline improves on sparse features by learning denser document representations, yet it still compresses the judgment too early. Once the document is collapsed into a single dense vector without strong contextualization, it becomes harder to preserve how different legal cues interact across the document. The gap between Doc2Vec and the proposed model therefore helps explain why contextual and sequence-aware modeling remain important in judicial corpora even when static distributed representations are available.

The randomly initialized BiLSTM occupies an intermediate position. It captures sequential patterns more effectively than classical baselines, but it lacks the legal-domain semantic priors supplied by transformer pretraining. Its results indicate that sequential modeling is beneficial, yet insufficient when domain-sensitive lexical interpretation must also be learned from limited supervision. This is precisely why the hybrid approach matters: one component supplies contextual legal semantics, while the other refines how those cues unfold across the document.

The stronger adapted neural baselines reinforce the same point. A LegalBERT-only classifier benefits from domain-specific pretraining and already provides a strong benchmark. The gains obtained by HyLegalML over that baseline indicate that the contribution of the proposed framework is not merely the use of a legal-domain transformer. It lies in the way contextual encoding, sequence refinement, calibrated thresholding, and conservative self-training are combined within one operational pipeline.

1. Deployment-oriented interpretation of the framework

A useful way to understand the framework is to ask how it would be used in a real legal information system. One deployment mode is automatic pre-tagging of newly ingested judgments before they are indexed in a repository. In this setting, the classifier does not need to be treated as infallible. Its value comes from reducing manual effort and creating a first-pass issue profile that can later be reviewed, corrected, or enriched by experts. The combination of Micro-F1 and LRAP reported in the study suggests that such pre-tagging is a realistic use case.

A second deployment mode is human-in-the-loop search support. A researcher searching for cases in a niche area may begin with a broad doctrinal category, then use predicted secondary labels to narrow or reorder the result set. In such a workflow, exact full-label recovery is desirable but not mandatory. What matters is that plausible secondary domains appear high enough in the ranking to attract attention. This is one reason the inclusion of LRAP is valuable: it speaks directly to suggestion-oriented use rather than only to hard classification.

A third deployment mode is corpus management for downstream machine learning. Balanced sampling for citation recommendation, rhetorical role labeling, or case similarity experiments often depends on being able to identify judgments from multiple legal areas efficiently. The proposed framework supports that need by providing a structured label space that can be used to build more representative subsets of a larger legal corpus. In this sense, the classifier is an infrastructural component for broader legal analytics, not merely an isolated endpoint task.

From a systems perspective, the framework is also attractive because it avoids dependence on unusually large hardware resources. A legal research group or academic lab can reasonably train and apply the model without access to industrial-scale infrastructure. That accessibility matters in practice. A slightly more accurate but far more expensive system may be less useful than a robust, well-understood, and reproducible model that can be adopted by a wider research community.

TABLE 10 Representative deployment modes for multi-label legal judgment classification

Deployment mode	Primary output use	Most relevant metric emphasis	Human oversight expectation
Repository pre-tagging	Initial doctrinal indexing of new judgments	Micro-F1 and Macro-F1	Post-hoc review for ambiguous or high-value cases
Search support	Ranking or filtering candidate judgments by likely secondary domains	LRAP and Macro-F1	Researcher examines suggestions before relying on them
Corpus construction	Balanced sampling for downstream legal NLP tasks	Macro-F1 and label coverage	Dataset curator validates domain balance and edge cases

J. Alternative long-document strategies and why they still matter

The present study intentionally adopts a resource-aware approach to long judicial documents, but this does not imply that the problem is fully solved. Long-context transformers remain an important alternative because they can preserve substantially more of the source judgment [16]. Their main attraction is that they reduce the need to decide which part of the document should be retained before encoding. This is particularly useful when legally significant reasoning appears in the middle of the judgment rather than near the beginning or end.

At the same time, long-context encoders come with real trade-offs. They are more computationally demanding, may require smaller batch sizes, and often complicate training stability in already low-resource settings. For a first journal version grounded in a practical judicial corpus, the Head plus Tail strategy is therefore reasonable. It enables a strong baseline system while leaving room for later work to examine whether longer-context architectures produce gains large enough to justify their cost.

Section-aware encoders form another promising middle ground. Instead of treating the judgment as one flat token sequence, future models could explicitly isolate fact sections, precedent discussion, statutory analysis, and holdings before applying domain prediction. Such an approach may be particularly helpful for minority secondary labels that are mentioned only in one rhetorically concentrated segment. The error patterns discussed earlier suggest that this sort of structural selectivity could be more valuable than merely expanding sequence length.

These alternatives underline an important point: the present framework should be understood as a strong, practical reference point rather than as the final word on judicial long-document classification. By establishing a solid performance level under realistic resource and annotation constraints, the study creates a defensible baseline against which richer long-context or structure-aware systems can later be judged.

K. Positioning the study within broader legal AI research

The present study occupies a specific position within the broader landscape of legal AI. It does not attempt to predict judicial outcomes, generate legal arguments, or replace doctrinal reasoning. Its contribution is more foundational: it seeks to improve the organization of judicial corpora so that later stages of analysis can operate on better-

structured information. This distinction is important because many ambitious legal AI systems silently depend on simpler upstream tasks such as classification, tagging, or metadata recovery. Without reliable document organization, even sophisticated downstream models may retrieve the wrong cases or learn from poorly balanced training sets.

This organizational role also helps explain why the proposed framework is worthwhile even though legal NLP has moved rapidly toward very large language models. General-purpose models can provide flexible zero-shot or few-shot predictions, but their outputs may vary across prompts, may be harder to calibrate for fixed taxonomies, and may impose higher inference costs for large corpus processing. A supervised or semi-supervised classifier built around a stable label set remains attractive when the goal is consistent large-scale indexing rather than open-ended generation. The present work therefore complements, rather than competes with, broader generative legal AI research. Another point of positioning concerns evaluation culture. Legal AI papers often emphasize a single benchmark score or compare systems that differ simultaneously in data source, label set, and preprocessing. The current study instead places emphasis on analytical separation: the corpus construction strategy is described explicitly, the benchmark is held out from silver-standard training, and multiple metrics are used to interpret utility from different operational viewpoints. This kind of staged evaluation is valuable in legal research because it makes the source of improvement easier to understand and easier to scrutinize.

Seen in this way, HyLegalML is best regarded as a carefully engineered baseline for Indian judicial domain tagging. It is strong enough to support legal analytics workflows, explicit enough to be reproduced and critiqued, and sufficiently modular to be extended by later work. That combination of practicality, transparency, and extensibility is often more useful for a field than a narrowly optimized model whose gains depend on opaque data preparation or aggressive benchmark tuning.

L. Immediate next steps for strengthening the journal version

The journal version of the work can be strengthened further in several focused ways. One improvement would be a fuller manual discussion of borderline cases in which expert annotators might reasonably disagree on whether a secondary domain should be assigned. Including such examples would help readers understand that some apparent errors arise from genuine doctrinal ambiguity rather than from careless prediction. A second improvement would be the addition of more explicit corpus-release or annotation-guideline details where feasible, since legal NLP reviewers increasingly expect dataset transparency alongside model results.

A further enhancement would involve direct comparison with at least one longer-context encoder under the same dataset conditions, even if that comparison is limited in scope. Such an experiment would clarify whether the remaining gap is primarily due to missing mid-document context or to broader issues of label overlap and data noise. Even when full large-scale comparison is not possible, a carefully framed pilot experiment could strengthen the methodological narrative of the paper.

Finally, the journal version would benefit from one or two illustrative examples showing how predicted labels correspond to realistic judicial issue profiles. Such examples need not be extensive, but they would make the discussion more tangible for readers who are less familiar with Indian legal documents. Combined with the quantitative results, these additions would further reinforce the central claim of the paper: that judicial domain classification is both feasible and useful when model design, corpus design, and evaluation design are treated as interconnected parts of the same research problem.

TABLE 11 Priority directions for extending HyLegalML in future journal revisions

Extension area	Expected benefit	Primary challenge
Longer-context encoders	Better recovery of mid-document cues and secondary labels	Higher memory cost and slower experimentation
Section-aware modeling	More precise use of facts, precedent, analysis, and holding segments	Requires reliable structural segmentation of judgments
Expanded court coverage	Improved external validity across Indian judicial tiers	Greater variation in formatting, language, and document quality
Illustrative case analysis	Stronger interpretability for reviewers and readers	Needs careful selection of representative examples without over-claiming

VIII. CONCLUSION

This paper presented HyLegalML, a hybrid transformer framework for multi-label classification of Indian legal judgments. The framework was designed for long documents, overlapping legal domains, class imbalance, and limited expert annotation. It combined LegalBERT-based contextual encoding with stacked BiLSTM sequence modeling, class-weighted training, label-specific threshold calibration, and pseudo-label guided semi-supervised refinement.

Experiments on a corpus of Supreme Court judgments showed that the proposed system outperformed both traditional and neural baselines across Micro-F1, Macro-F1, Subset Accuracy, and LRAP. The gains were supported by ablation analysis, which demonstrated that each of the major components contributed independently to the final performance. The results indicate that multi-label legal classification in the Indian judicial setting benefits from a combination of contextual semantics, order-sensitive modeling, and conservative semi-supervised expansion.

The study provides a practical basis for domain tagging in legal information systems and can support downstream tasks such as legal retrieval, citation recommendation, and corpus analytics. Future work may extend the approach to broader Indian court corpora, richer hierarchies of legal domains, and long-context or citation-aware classification architectures.

Funding Statement: The author received no specific funding for this study.

Conflicts of Interest: The author declares that there are no conflicts of interest to report regarding the present study.

REFERENCES

- [1] Y. Yang and X. Liu, "A re-examination of text categorization methods," in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, pp. 42-49.
- [2] G. Tsoumakas and I. Katakis, "Multi-label classification: an overview," International Journal of Data Warehousing and Mining, vol. 3, no. 3, pp. 1-13, 2007.
- [3] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in Proceedings of NAACL-HLT 2016, 2016, pp. 1480-1489.
- [4] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "Large-scale multi-label text classification on EU legislation," in Proceedings of ACL 2019, 2019, pp. 6314-6322.

- [5] I. Chalkidis, E. Manavopoulos, M. Fergadiotis, and I. Androutsopoulos, "Extreme multi-label legal text classification," in *Proceedings of the Natural Legal Language Processing Workshop 2019*, 2019, pp. 78-87.
- [6] D. Tuggenier, P. Von Däniken, C. Peetz, and T. Cieliebak, "LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts," in *Proceedings of LREC 2020*, 2020, pp. 1235-1241.
- [7] 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.
- [8] V. Malik, R. Sanjay, S. Nigam, K. Ghosh, and A. Bhattacharya, "ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation," in *Proceedings of ACL-IJCNLP 2021*, 2021, pp. 4046-4062.
- [9] I. Chalkidis, M. Jana, M. Hartung, N. Bommarito, D. Androutsopoulos, D. Katz, and N. Aletras, "LexGLUE: A benchmark dataset for legal language understanding in English," in *Proceedings of ACL 2022*, 2022, pp. 4310-4330.
- [10] A. Kapoor, M. Dhawan, A. Goel, A. T. H. Arjun, A. Bhatnagar, V. Agrawal, A. Agrawal, A. Bhattacharya, P. Kumaraguru, and A. Modi, "HLDC: Hindi legal documents corpus," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 3521-3536.
- [11] I. Nair and N. Modani, "Exploiting language characteristics for legal domain-specific language model pretraining," in *Findings of the Association for Computational Linguistics: EACL 2023*, 2023.
- [12] A. Joshi, S. Paul, A. Sharma, P. Goyal, S. Ghosh, and A. Modi, "IL-TUR: Benchmark for Indian legal text understanding and reasoning," in *Proceedings of ACL 2024*, 2024, pp. 11460-11476.
- [13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT 2019*, 2019, pp. 4171-4186.
- [14] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of ICML 2014*, 2014, pp. 1188-1196.
- [15] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine Learning*, vol. 39, no. 2-3, pp. 103-134, 2000.
- [16] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.
- [17] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "RCV1: A new benchmark collection for text categorization research," *Journal of Machine Learning Research*, vol. 5, pp. 361-397, 2004.