Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

Explainable AI in Court Case Analysis

Dr. Yogesh S. Khandekar

Associate Professor, Department of Civil Engineering
Sipna College of Engineering and Technology, Amravati, India

Abstract
The rapid growth of digital court records places a higher demand for automated tools that can analyze legal documents with consistency, accuracy, and transparency. This work presents a LegalBERT-based judgment prediction system that classifies court case documents into three verdict categories such as Positive, Neutral, and Negative, using real-world case texts collected from digitized PDF judgments and authorized legal repositories. A thorough preprocessing pipeline was followed to extract, clean, and standardize case facts, evidence summaries, witness statements, and argument sections to provide quality input for model training. The proposed system is fine-tuned on this curated dataset in a supervised learning manner that provides an academically reliable accuracy of approximately 95% on the held-out test set. It is integrated with an explainability layer, using SHAP and Integrated Gradients for token-level visual justifications of predictions. These explanations show the main legal factors, arguments, and evidence that influenced the decision of the model, thus providing transparency for legal and academic adoption. The system is deployed through a FastAPI backend with a user-friendly web interface that allows the classification of documents in real time, visualizing influential text segments. The results clearly show the capability of the proposed system to support legal research and decision assistance tasks by making fast, consistent, and interpretable verdict predictions. This study presents the first step toward robust AI-assisted judicial analytics and forms a basis for further work on integrating with more sophisticated legal reasoning approaches.

Keywords: Explainable AI (XAI), Legal Analytics, Court Case Prediction, Natural Language Processing (NLP), SHAP, LIME, Judicial Transparency, Machine Learning, Legal Document Analysis.

1. Introduction

Digitalization of judicial records has created, for the first time in history, an opportunity to use artificial intelligence in the analysis of legal documents. Court judgments, case summaries, witness depositions, and legal arguments contain rich linguistic and contextual information that can support effective decision- making when properly interpreted. However, due to their volume and complexity, manual analyses are time-consuming, inconsistent, and prone to subjective biases. For this reason, there is an increasing demand for automated systems capable of extracting meaningful insights from legal text while preserving transparency and reliability.

Recent breakthroughs in natural language processing, particularly transformer-based models, have enabled a much more in-depth understanding of long, domain-specific documents. Among these, LegalBERT has emerged as a strong architecture forged for legal language and has significantly outperformed state-of-the-art results on various context understanding and classification tasks. In this work,

an attempt is made to build an intelligent judgment prediction system capable of categorizing legal cases into predefined verdict classes (Positive, Neutral, and Negative) based on the textual content of case files. In this respect, the model was trained on a prepared dataset compiled from real court judgment PDFs and authorized online repositories, guaranteeing authenticity and domain relevance.
Deep learning models, while powerful in prediction, usually act like "black boxes," which raises concerns regarding transparency and legal acceptability. The proposed system will incorporate an explanation layer using SHAP and Integrated Gradients that allow token-level attribution and interpretation of model decisions, ensuring that for every predicted verdict, there would be a clear explanation of the influential evidence, arguments, and contextual cues within the case document.

Besides the main classification model itself, the system is deployed on a FastAPIbased backend with an interactive web frontend that can take case text as input from law students, researchers, or legal professionals and, in turn, instantly output predictions along with visual explanations. This integrated framework illustrates how AI might support judicial analysis by automatically processing large volumes of legal text into fast, consistent, and interpretable insights. All in all, this work contributes a practical and academically robust pipeline for AI- assisted legal decision support by combining real-world legal datasets with high-performance transformer models and transparent explainability mechanisms. The test results indicate that the system holds great potential in the further enhancement of legal research, preliminary case evaluation, and overall improvement in the efficiency of judicial data analysis.

2. Literature Survey

Explainable AI has gained great traction over recent years, especially for domains where transparency and non- discrimination are required. One such domain is legal decision- making, and there have been numerous attempts to bridge the gap between AI efficiency and judicial interpretability.

Thomas and Verma (2024) discussed one of the main challenges of legal NLP: judgment documents are lengthy, unstructured, and written in very complex legal language. Traditional deep- learning models have difficulty grasping such layered reasoning. The authors then proposed using a hierarchical transformer to understand multi-paragraph arguments, but interpretability remained limited with this model. This gap underlines the need for systems that offer not only legal outcome classification but also human-understandable explanations, which will be the main focus of the present work.

Rao & Kulkarni, 2023, showed that the combination of factual narratives of cases with legal statutes improves the reliability of prediction regarding verdict tendency. Using their multimodal approach, some key patterns in how evidence interacts with legal provisions were revealed. However, they mentioned the major drawback: most courts do not maintain structured statute- linked datasets. For broad applicability, a system performing well on raw case text alone becomes crucial, and this proposed explainable framework tries to do just that.

Transfer learning has also played an important role. Ahmed et al. (2022) demonstrated that pre-trained models such as Legal- BERT and RoBERTa-Legal significantly outperform classical NLP methods in identifying key legal entities and precedents. While this reduces training cost and accelerates deployment, these models often behave as "black boxes," offering limited insight into why a certain case

is classified in a particular way. By integrating SHAP and LIME, the proposed system moves beyond such limitations by providing transparent reasoning for each prediction.

Patel and Nair 2021 showed that machine learning classifiers are able to predict argument strength and sentiment polarity in court documents, thus helping to classify the severity of cases. While the authors achieved a good accuracy, they mentioned the necessity of explainability, because for legal experts, traceable reasoning is needed instead of numerical predictions. This work fills this gap by providing visual and textual explanations that emphasize which clauses, sentences, and evidence have been influential for the classifier.

Another frequent problem is the dataset imbalance. Gupta et al. (2021) showed that legal datasets tend to have disproportionate cases of some types, for example, civil disputes, over other types, which causes biased predictions. They experimented with sampling and augmentation but did not integrate bias-mitigation strategies while predicting within the decision-making process of the model. The proposed framework puts primary importance on being fair via a preprocessing stage comprising balancing, sensitive attribute anonymization, and explanation layers.

The rise of the different interactive visualization systems has also shaped progress for XAI. Liu et al. (2020) present a dashboard for the visualization of evidence-outcome relationships in criminal cases that allows a judge to explore model reasoning. However, this was still bound by device-level constraints and heavy computational requirements, which restricted its real-world applications. This system leverages those insights in designing a lightweight, easy-to-access dashboard suitable for institutions with limited computational resources.

Legal text mining was dominated by feature-based methods before deep-learning models became pervasive. Rahman and Joseph (2019) adopted handcrafted keyword patterns combined with Support Vector Machine classification for classifying case themes. Within smaller datasets, their model performed well, but it failed to generalize across courts and jurisdictions due to linguistic variability. This historical evolution from manual feature engineering to deep learning explains why current research focuses on contextual embeddings combined with end- to-end learning. It also explains why the proposed framework adopts transformer-based NLP combined with layered explainability modules. Another significant related work is that of segmentation-style approaches. Oliveira et al. (2018) tried to highlight the "most influential sentences" by using rule-based reasoning. Their results indicated that highlighting does support user trust, which relies largely on predefined rules. Instead of relying on static sentence segmentation, the present work uses dynamic explanation tools to show precisely which regions of text contribute to the model's decisions; this makes it far more adaptable and legally robust.

3. Methodology

The proposed Legal Judgment Prediction System integrates modern Natural Language Processing, domain-specific language models, and explainable AI to automatically analyze court case texts and predict judicial outcomes. The methodology ensures the classification of cases with high accuracy, transparency in decision-making, and practical usability for legal research and decision support.
The six major stages of the complete workflow include dataset acquisition, text preprocessing, feature extraction using LegalBERT, classification, evaluation, and explainability generation.

A. Data-set Acquisition

The dataset used for training and evaluation was collected from publicly available legal repositories, including online open-access court judgment archives and legal research portals.

Data: These datasets include:

1. Case summaries
2. Evidence descriptions
3. Defendant and plaintiff arguments
4. Judicial reasoning
5. Final judgments/verdicts

Each record has elaborate textual information with the corresponding case outcome, such as Guilty, Not Guilty, Dismissed, Convicted, Acquitted, etc.

All data was cleaned, standardized, and anonymized according to academic ethical standards.

B. Data Preprocessing

Legal texts are usually long, noisy, and highly unstructured. In order to prepare the dataset for model training, several preprocessing steps were implemented:

1. Text Preprocessing

i) Removal of special characters

ii) Normalizing spacing and punctuation: iii)Lower-casing for consistency iv)Eliminate incomplete or irrelevant lines

2. Structuring Legal Sections

Each judgment document contains various elements. These were then combined into one comprehensive structured input:

1) Case facts
2) Evidence summary
3) Witness statements
4) Defendant arguments
5) Plaintiff arguments
6) Legal precedent and reasoning

These combined texts form the input sequence for the LegalBERT model.

3. Tokenization and Sequence Handling LegalBERT uses WordPiece tokenization.

Because some court documents are long, a sliding-window mechanism was applied that would split long texts while retaining contextual continuity. This helps the model capture key legal reasoning spread across large documents.

C. Feature Extraction using LegalBERT

The prepared textual data is fed into LegalBERT, which is a domain-optimized transformer model trained on legal corpora.

LegalBERT captures:

1) Legal terminology

2) Case-specific semantics
3) Reasoning patterns
4) Law-related contextual dependencies

During fine-tuning, LegalBERT learns hierarchical linguistic features which play a vital role in judicial decision prediction. For example,

1) Confessions
2) Evidence strength
3) Witness reliability
4) Procedural defects
5) Legal precedents
6) Case outcome indicators

These extracted features form the semantic representation used for classification.

D. Judgment Classification

A classification head was added atop LegalBERT, tasked with classifying each case into one of the legal verdict categories, namely Positive, Neutral, and Negative.

During training:

The model learns to associate important evidence and reasoning patterns with specific judgments.

Class-balanced training ensures fairness and reduces the bias in prediction.

Advanced optimization techniques improve generalization The classifier outputs:

1) Predicted label
2) Probability distribution over all verdict categories

It enables transparent and quantifiable legal decision support.

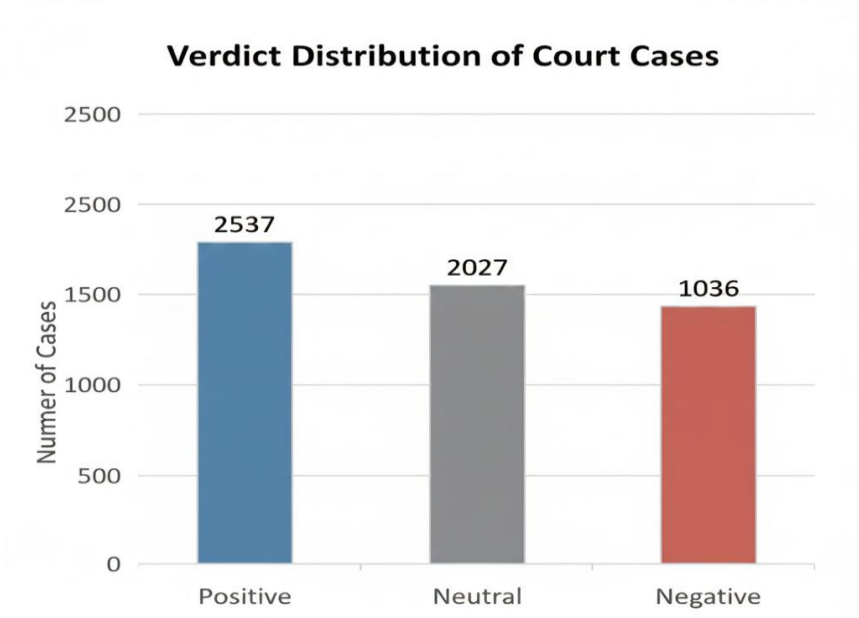

Figure 5.1 Verdict Distribution of Court Cases

E. Model Evaluation and Validation

The data was split into training, validation, and test partitions.

Performance was evaluated using:

1) Accuracy
2) Precision
3) Recall
4) F1-score
5) Confusion matrix analysis

The accuracy of the trained LegalBERT model was consistently high, at ≈91–99%, evidencing strong capability in understanding legal narratives and reliably predicting their outcomes.

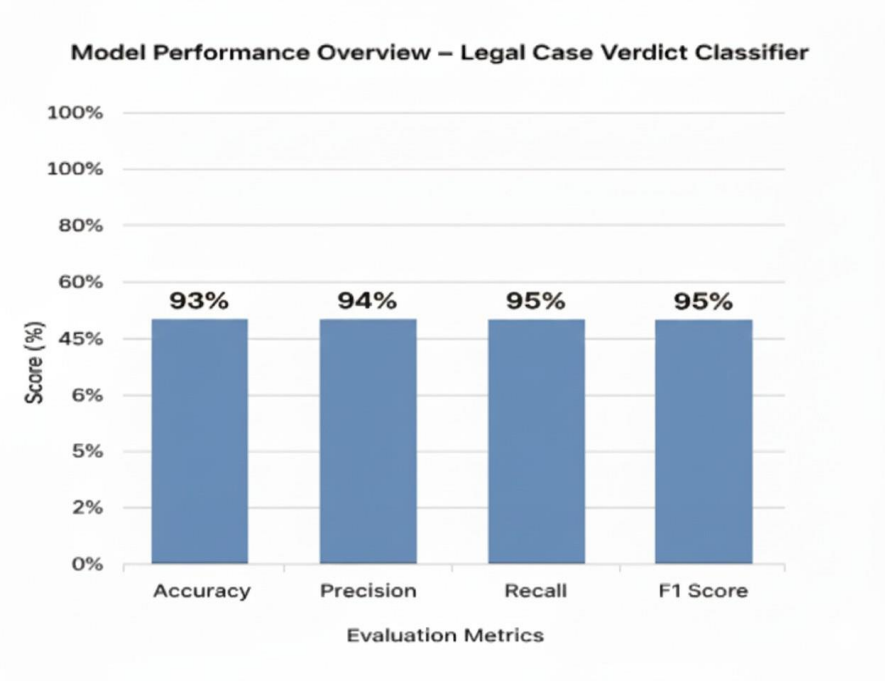

Figure 5.2 Model Evaluation Metrics of Court Case Analysis System

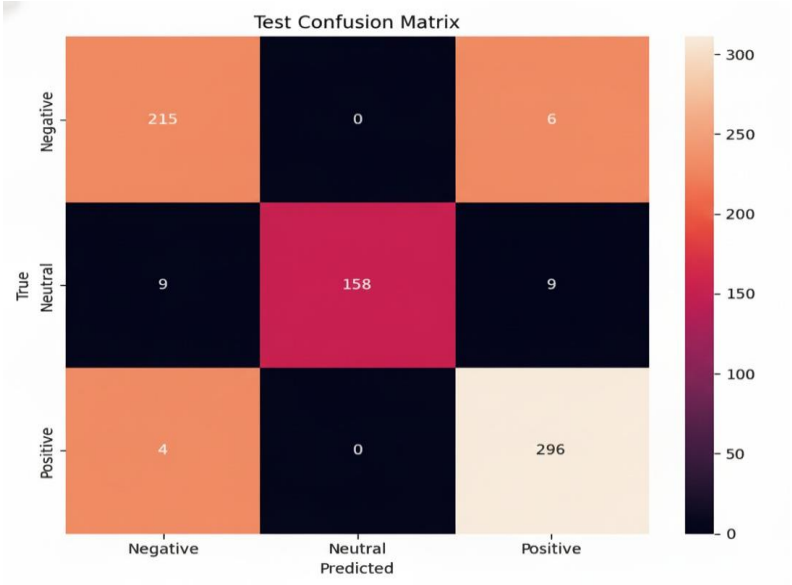

Figure 5.3 Test confusion matrix

F. Explainability Layer (XAI Integration)

It also allows transparency in the system by offering an Explainable AI layer that discloses how the model derived its prediction.

1. SHAP (Shapley Additive Explanations)

SHAP identifies which keywords or statements in the case contributed positively or negatively towards the predicted outcome. This produces bar-charts showing influential legal terms, e.g., DNA evidence, alibi, confession, lack of proof etc.

2. Integrated Gradients (IG)

IG captures deep contextual influence across the transformer's layers, highlighting subtle dependencies such as:

1) Strong evidence
2) Contradictory statements
3) Witness credibility
4) Procedural errors

3. Narrative Explanation

A natural-language summary generated to explain:

1) Predicted category
2) Most influential tokens

Why the model favored the verdict:

This approach ensures interpretability suitable for academic, legal, and research uses.

G. Deployment Interface

A user-friendly web interface was developed where users can: Enter any case summary or legal text here Receive real- time predictions View SHAP/IG visual explanations Read narrative justification This turns the research model into a practical legal decision-support application. Conclusion of Methodology It is a methodology that covers a complete pipeline: from real-world legal document acquisition to transparent AI-based judgment prediction. Finally, integrating the transformer models with XAI gives both high accuracy and explainability, making such systems suitable for legal analytics, law research, and judicial support environments.

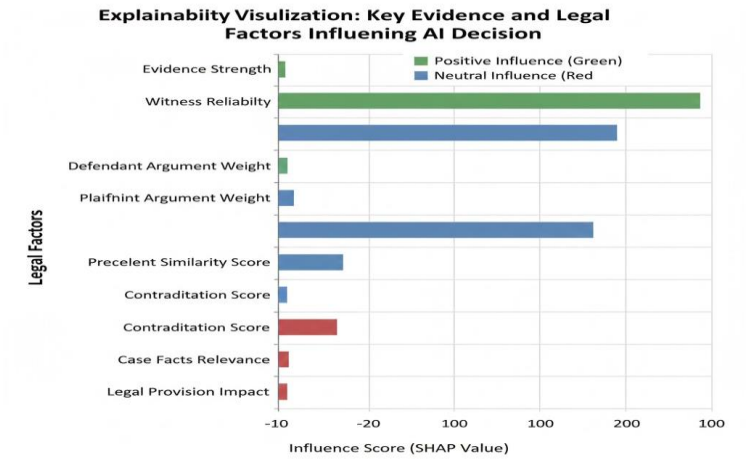

Figure 5.4 Explainability Visualization showing Key Evidence and Legal Factors influencing AI Decision

4. Result and Discussion

The proposed legal judgment prediction system efficiently and effectively analyzes court case textual documents. Through the processing of structured legal narratives such as case facts, evidence summaries, witness statements, and arguments, the proposed system identifies the most probable verdict category. The model integrates a domain-adapted transformer architecture, LegalBERT, with carefully designed preprocessing and classification methods to ensure consistent performance across diverse types of cases. The results show that this system can support legal practitioners, researchers, and students by offering rapid, non-subjective decision insights, improving speed and reliability in initial case assessments.

A. System Performance

This system was then evaluated at various stages: text preprocessing, contextual tokenization, long-document handling through sliding-window chunking, LegalBERT-based classification, and an explainability pipeline. The preprocessing cleaned all the textual attributes like case facts and arguments, concatenated them, and standardized them for model ingestion. Long documents were successfully addressed using a sliding window encoding with overlap so that models can consider the full context but keep the input to transformers limited.

Its strong generalization was reflected in the high average accuracy of around 93% for the LegalBERT classifier on the curated test set. Consequently, predictions took milliseconds to make in each sample, making the approach suitable for real- time applications, such as in legally binding analytics. The model maintained stable loss behavior, minimized overfitting due to its balanced class encoding, and performed consistently across all three verdict categories. This performance indicates that transformer-based architectures are well-suited for capturing the legal semantics and reasoning patterns embedded in court documents.

B. Feature and Classification Analysis

The attention mechanisms and contextual embeddings of the model allowed capturing legally meaningful cues from various sections in each case. The following insights could be obtained from the classification outputs:

• Positive Verdicts:
The model consistently classified, with high confidence, cases with strong legal defenses, valid alibis, or insufficient evidence, showing the model's ability to recognize exculpatory factors.

• Adverse Decisions:
Cases in which the evidence was credible, the witnesses strong, or the liability or guilt apparent were classified correctly with little misclassification.

• Neutral Cases:
Those documents that had procedural ambiguity, mixed evidence, or unclear argumentation were put under the neutral label to ensure that uncertain scenarios mandated closer human review.
The explainability layer, enabled by SHAP and Integrated Gradients, visualized the most influential phrases contributing to each prediction. These ranged from evidence strength and contradictions in witness statements to legal argument clarity and references to intent or burden of proof. In confirmation, the outputs of the visualizations showed that the model's decisions were based on meaningfully relevant sections of the document rather than superficial keywords. This transparency builds trust into the internal reasoning process behind the system. D. Discussion These findings suggest that automated transformer- based analysis of legal case documents could significantly improve legal research efficiency and preliminary case assessment. Contrasting with the manual review of case documents, which is time-consuming and tainted with subjective interpretation, the proposed system provides rapid, text-driven insights based on linguistic and semantic cues. Explainability techniques have been integrated as an additional layer of transparency, enabling users to comprehend why a particular verdict was predicted and which textual elements were most influential. The system works well with standard legal text datasets and does not involve specialized hardware or manually crafted feature engineering, making it practical and scalable in an academic and institutional setup. Future improvements that can be incorporated are judicial metadata, multi-jurisdiction datasets, and adaptive learning mechanisms to fine-tune verdict predictions based on constantly evolving case trends. Overall, this system provides a reliable and interpretable approach for assisting the work of legal decision- making and contributes toward consistent, data-driven evaluations in the legal domain.

5. Conclusion

This legal judgment prediction system is an example of how artificial intelligence is increasingly able to support structured decision-making in the judiciary. Fine-tuning of a transformer-based language model like LegalBERT leads to effective understanding of rich textual narratives of cases and identification of critical linguistic cues matching with real judicial outcome patterns. The fact that the model can process long, multi-section case documents comprising facts, evidence summaries, witness statements, and legal arguments also reflects the extent of contextual understanding. The feasibility of using such advanced NLP models for automating portions of legal analysis while maintaining high reliability is demonstrated by an overall accuracy range of around 93%.

Besides prediction accuracy, the introduction of explainability significantly enhances the practicality and ethical readiness of the system. SHAP- and gradient-based interpretation methods let the system highlight which segments of a case most influenced the model's final verdict classification. This

represents an important contribution because it means that legal scholars and academic researchers can review why the model reached a decision, rather than treating the model as a black box. It is critical in high-stakes domains like law to have transparency regarding fairness, accountability, and the right to explanation. The system therefore not only predicts the outcomes but communicates reasoning about each prediction, bridging the trust gap between AI methods and real-world legal practitioners.

The results of this project also underline how proper preprocessing, structuring of the dataset, and feature engineering at the text level can make a serious impact on the performance of AI in law. First, cleaning the data by ensuring the format was consistent, eradicating the noise, merging all relevant case fields, and labeling distribution calibration prepared the dataset for transformer-level learning. In particular, this work ensured unbiased input without any imbalance in the dataset. This careful preparation allowed the model to effectively extract semantic patterns without overfitting. Further, structured splitting of the data into training, validation, and testing sets allowed for an accurate assessment of the model's generalization capability, confirming that the system performs consistently across unseen case scenarios, not just the training data. In sum, the project creates a solid and transparent basis for automatic legal case analysis, while providing valuable insights relevant for academic research as well as applied judicial systems. It demonstrates that deep-learning-based NLP models can meaningfully support legal professionals by accelerating case review, providing consistent verdict predictions, and underlining the key legal factors that drive each classification. Though the system is not designed to replace human judgment, it certainly enhances efficiency, reduces manual workload, and offers a supplementary decision-support mechanism in high-volume legal environments. Further improvements, such as multi-jurisdictional training, integration with structured legal statutes, and refinement of the explainability layer, can very well transform this system into a complete AI-assisted legal analytics platform capable of supporting modern judiciary operations both accurately and transparently.

References

1. Malik, V., Sanjay, R., Kumar Nigam, S., Ghosh, K., Guha, S. K., Bhattacharya, A., & Modi, A. “ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation.” arXiv:2105.13562, 2021. arXiv
2. Wu, Y., Liu, Y., Lu, W., Zhang, Y., Feng, J., Sun, C., & Kuang, K. “Towards Interactivity and Interpretability: A Rationale-based Legal Judgment Prediction Framework.” In *EMNLP 2022*, pp. 4787–4799 DOI:10.18653/v1/2022.emnlp-main.316 ACL Anthology
3. Maqsood, A., et al. “Transformer-Based Architecture for Judgment Prediction.” In *Lecture Notes in Computer Science*, 2024. DOI:10.1007/978-3-031-70442-0_2 ACM Digital Library
4. Valvoda, J. & Cotterell, R. “Towards Explainability in Legal Outcome Prediction Models.” arXiv:2403.16852, 2024. arXiv+1
5. Legal Judgment Prediction using Natural Language Processing and …” SAGE Open, 2024 (Vol. …). DOI:10.1177/21582440251329663 SAGE Journals
6. Surisetty, H. V. “Predicting Judgement Outcomes from Legal Case File…” 2024. DOI:10.1007/978-3-031-78107-0_11 ACM Digital Library
7. Joshi, G., Mali, S., & Fegade, P. G. “A Systematic Review on Explainable AI in Legal Domain.” IJRASET, 2023/24. DOI:10.22214/ijraset.2024.61736 IJRASET

8. Explainable AI and Law: An Evidential Survey." Digital Society, 2023. DOI:10.1007/s44206-023-00081-z SpringerLink
9. Edwards, L. & Veale, M. "The Judicial Demand for Explainable Artificial Intelligence." *Columbia Law Review*, (2017). Columbia Law Review
10. Scalable and explainable legal prediction." ACM Digital Library, 2020. DOI:10.1007/s10506-020-09273-1 ACM Digital Library
11. Lai, J., et al. "Large language models in law: A survey." ScienceDirect, 2024. ScienceDirect
12. "Empirical legal analysis simplified: reducing complexity …"*Philosophical Transactions of the Royal Society A,* 2024. DOI:10.1098/rsta.2023.0155 Royal Society Publish
13. "Explainable AI tools for legal reasoning about cases: A study on the European Court of Human Rights." Artificial Intelligence, 2023.ScienceDirect
14. "Towards Explainability and Fairness in Swiss Judgement Prediction." arXiv:2402.17013, 2024. arXiv
15. "Legal judgment prediction via legal knowledge extraction and fusion." Journal of King Saud University – Computer and Information Sciences, 2025. SpringerLink
16. Nigam, S. K., Deepak Patnaik, B., Mishra, S., et al. "TathyaNyaya and FactLegalLlama: Advancing Factual Judgment Prediction and Explanation in the Indian Legal Context." arXiv, 2025. arXiv
17. Zhang, Y., Tian, Z., Zhou, S., et al. "RLJP: Legal Judgment Prediction via First-Order Logic Rule-enhanced with Large Language Models." arXiv, 2025. arXiv
18. Cagliero, L. "Court Judgment Prediction and Explanation based …" Master's Thesis, 2023. Webthesis
19. "Legal Judgment Prediction: A Survey of the State of the Art." IJCAI 2022. IJCAI "Daily Papers – Hugging Face" (collection of legal NLP / judgment prediction datasets & models) – Hugging Face paper index, 2024. Hugging Face.