International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 04 | Apr 2026 www.irjet.net p-ISSN: 2395-0072

AI Driven Legal Document Intelligence

Shaik Riyaz [1], Abhishek Arun Kumar [2], S. Harshitha [3], P. Varshita[4]

[1]Shaik Riyaz, Senior Assistant Professor Department of Computer Science and Engineering, Geethanjali College of Engineering and Technology Hyderabad, India
[2]Abhishek Arun Kumar, Student, Department of Computer Science and Engineering, Geethanjali College of Engineering and Technology Hyderabad, India
[3]S. Harshitha, Student, Department of Computer Science and Engineering, Geethanjali College of Engineering and Technology Hyderabad, India
[4]P. Varshita, Student, Department of Computer Science and Engineering, Geethanjali College of Engineering and Technology Hyderabad, India

---

Abstract - *The introduction of Artificial Intelligence in the legal profession has created significant changes in terms of document processing, analyzing, and management. In this regard, the research will explore the concept of "AI Driven Legal Document Intelligence," concentrating on the use of AI in contract review, legal analysis, and legal document data extraction. Specifically, it is expected that through the use of AI, including the implementation of algorithms based on Natural Language Processing and Machine Learning, it would be possible to considerably shorten the process of analyzing a large number of documents while enhancing the detection of risks and obligations. Moreover, the transition from conventional keyword searching to the use of semantic analysis and gaining will be analyzed based on the extent to which better understanding of the document's contents is achieved.*

Key Words: **Artificial Intelligence, Legal Technology, Document Intelligence, Natural Language Processing, Machine Learning.**

1. INTRODUCTION

An AI-Driven Legal Document Intelligence System project entails creating an automated system that will be able to analyze legal documents including contracts, agreements, legislation, and case laws. In practice, experts have to manually go through complex legal documents to get meaningful information out of them, but the process takes a lot of time and may contain errors committed by humans. In the proposed project, modern Natural Language Processing tools will be used for processing legal texts using transformers like LegalBERT, CaseLaw-BERT, and Longformer. Among other functions performed by this system are clause extraction, risk prediction, and document summarization, making it possible to provide quality services regarding legal document analysis. Additionally, methods like SHAP and LIME will be applied to interpret algorithms of AI in a clause-by-clause manner. Apart from that, a web application with a dashboard will also be part of the project where risk prediction, important clauses, and document summaries will be presented visually.

There are a number of challenges associated with existing analysis systems of legal documents. The major problem faced by most of these systems is that they only have access to limited data sets, and therefore they cannot be used in analyzing different kinds of legal documents like contracts, case law, and statutes. Besides, most of the artificial intelligence approaches in use today act as black boxes where they produce prediction but fail to explain how the decision was arrived at. These limitations make it difficult for many users to adopt such systems since they are not transparent and accountable. Moreover, most of these systems do not integrate multiple sources of legal data, making the analysis incomplete and lacking in contextual understanding. Other challenges include lack of a practical interface and failure to be deployable in a real-world setting. As a result, there is a need for an advanced AI system that analyzes legal documents.

2. LITERATURE REVIEW

Significant advances have been achieved in the field of Artificial Intelligence (AI) and NLP, which means that now the analysis of legal documents is based not only on keyword search but also on semantics. Quite a few studies can be found that investigate machine learning and deep learning models aimed at clause extraction, finding liability, and ensuring compliance with regulations among a huge body of legal texts. Thus, this literature review constitutes a good starting point for creating an intelligent system that is going to be used for contractual risk assessment. The research work Automated Legal Risk Assessment (2024) employs CNN and BERT models that are meant to classify clauses and predict risks of litigation. The system performs extremely well in terms of risk prediction; however, it is designed exclusively for working with one document, and there is no feature of translation into other languages.

The research article entitled "Transformer-based Clause Analysis for Indian Law (2025)" uses LEGAL-BERT as an aid in accurately classifying contractual duties. It emphasizes the importance of AI technology in differentiating between regular language and aggressive language within the indemnity clause. Despite the study's emphasis on accuracy

in results, it operates in a black box manner, and there are no visualization techniques like SHAP and LIME to explain why certain clauses are identified as being aggressive. Furthermore, it does not offer any clear-cut guidelines and leaves the interpretation of findings up to the discretion of the reader. The other significant research article discussed in the literature review is the "Hybrid LSTM & LLM based model for Contract Summarization (2024)". This article demonstrates the ability of LLM's structure in capturing text dependencies. However, there is no persistence layer that can be used for handling data efficiently in the database.

Moreover, a review article published in Legal Analytics (2025) conducts a comparative analysis of different machine learning algorithms like SVM and Random Forests for contract auditing. While SVM is identified as one of the best models in this study, it overlooks the inclusion of advanced elements from Generative AI, risk dashboards, and prediction features necessary for real-world applications in law. Clearly, all previous models focus on the dichotomy of Risk versus Non-risk cases and lack a holistic model for multi-document ingestion (PDF/Docx), clear-cut "AI Verdicts," and explainable AI. In this context, LexAI: Legal Document Intelligence Platform aims to solve these issues by providing multiple sources of data mining, AI Verdict, and dashboards using SHAP/LIME approach.

3. EXISTING SYSTEM

At present, legal documents analysis and risks monitoring tools are usually standalone systems, which focus on a particular aspect of analysis, such as clause extraction or classification instead of combining all the tools within one unified analysis system. The first type of existing systems largely uses conventional machine learning and initial deep learning approaches to analyze the dataset or the case laws academically. Despite the usage of advanced transformers, for example, BERT, Legal-BERT, or CaseLaw-BERT, such systems can be defined as “single-purpose” solutions. In other words, they perform well when analyzing legal terms or detecting core clauses such as Indemnity, Confidentiality, or Termination. Nevertheless, the output is usually unstructured and unprocessed, meaning that the user has to interpret the results independently.

Another type of the existing systems is represented by customized architectures such as Longformer that can be used to analyze very lengthy documents including judicial transcripts and multi-page agreements. Even though such technologies allow processing a large number of tokens in just seconds, the functions of providing plain-language summaries, sentiment analysis or even jurisdiction comparisons are missing. As a result, legal experts have to spend a lot of time reading and analyzing enormous amounts of text before realizing what risks they have to face. Moreover, most of the existing solutions operate in isolation; for instance, the system may determine that a certain clause belongs to a "high-risk" category, but fails to draw final conclusions and to make the connection between the clause and its interpretation by common people. Thus, currently, there are no unified tools for combining heterogeneous perspectives on various legal aspects into one holistic view. The majority of the available platforms do not support such important functionalities as real-time risk assessment or identification of jurisdictional differences.

Although there have been significant improvements in artificial intelligence technologies within the field of law, certain fundamental weaknesses remain within the methodologies adopted thus far. One major limitation is the use of a single source database, whereby only one of two sources, such as the company’s contract agreements or case laws, is analyzed. The approach prevents the model from generalizing across the various legal jurisdictions present in an actual business environment. Furthermore, the failure to integrate information from the various sources leads to incomplete comprehension of the context, whereby the model understands the literal interpretation of the contract but fails to understand its risk implications.

In addition, the vast majority of such systems is based on the “black box” concept. They give accurate results but do not provide any explanation for these decisions, which leads to poor reliability due to the absence of rational arguments and justification (explainability gap). Furthermore, existing solutions are merely scientific tools that cannot be put into practice and used because of the lack of proper visual interfaces and intuitive dashboards. Not having such tools, users are unable to understand and interpret data easily. Lastly, modern systems face problems concerning the lack of real-time processing and actionable guidance because they generate plain data rather than recommendations such as “Safe to Proceed” or “Consult a Lawyer.” All these issues emphasize the importance of developing an all-around solution like the proposed LexAI system.

4. PROPOSED SYSTEM

The LexAI: Legal Document Intelligence Platform represents an all-encompassing and end-to-end solution aimed at streamlining legal analysis. The use of cutting-edge natural language processing techniques combined with a user-friendly design enables a clear-cut and actionable interpretation of any legal document.

A. Document Ingestion & Normalization Module
This is the initial stage where the system processes multiple input types such as PDFs, DOCX files, and TXT documents. The system applies machine learning algorithms to extract information from the legal document and normalize its structure, regardless of how unstructured and scanned it is.

B. Structural Clause Segmentation Module
In this phase, the system leverages semantic parsing techniques to divide long-form texts into separate clauses. This helps the system identify different parts of the text based on structural cues such as headings (Indemnity, Force Majeure).

C. Simple Language Translation Module
This module aims at translating complicated legal jargon and language into understandable and easy language so that ownership rights and responsibilities can be easily comprehended by non-professional people and small business owners.

D. Risk Assessment & Scoring Module
It is basically the analytic brain that uses a domain-specific transformer model like LEGAL-BERT to conduct assessment and assign a numerical risk rating from zero (0.0) to one (1.0) based on the comparison between the text and enterprise liability standards.

E. Verdict Engine
The decision-making brain of the system, which takes all the risk scores in the document to give a clear conclusion about the document as a whole in terms of its risks being low, high, or contacting a lawyer.

F. Explainable AI (XAI) Insight Module
To ensure trust among professionals, the interpretability features such as SHAP and LIME are incorporated. This allows for identification of the particular "trigger words" in the clause that lead to the red flag warning.

G. Interactive Visuals & Dashboard Module
The visualization module ensures the user has access to the high-fidelity interactive experience to explore the information provided. The user can utilize such features as risk distribution charts or side by side feed comparisons, which help to filter out and explore specific legal issues.

H. Export & Reporting Module in Multiple Formats
This module enables the creation of an executive summary report as well as the formal analysis report in either PDF or JSON format. The module is intended to be used in the enterprise setting where stakeholders will need to share their findings or record versioning history or risks.

Workflow Summary
As was mentioned above, the workflow starts with document ingestion and segmentation. Further, risk classification based on neural network takes place. These insights are interpreted by the XAI module, converted into human-readable format, and presented in visuals on the interactive dashboard. Finally, the decision is reached, and persistent data recorded.

5. METHODOLOGY

This is a systematic approach in developing models through iteration and incrementation designs. First, legal documents including contracts, laws, and statutes are gathered. These documents will be preprocessed by cleaning, tokenizing, normalizing, and segmenting them prior to being fed into the model. Legal datasets will be employed in order to train and test several transformer models, such as LegalBERT, CaseLaw-BERT, and Longformer. Models are evaluated based on accuracy, precision, recall, and F1-score, thereby determining which model is most efficient. The model includes an interpretability approach including SHAP and LIME. Finally, deployment will include uploading the document onto the web application via frameworks such as Flask.

6. SYSTEM ARCHITECTURE

For LexAI model, it was made in such a way that apart from encouraging modularity, it would also permit scalability and be able to handle legal textual data in complex ways. In each of the layers, there are unique processes performed by each layer, which collectively ensure the proper analysis of documents and identification of hidden legal threats. The system integrates various phases into one coherent process just like in any well-oriented judicial analysis.

1. Data Layer
The Data Layer serves as the foundation for the repository within the system, where it controls the process of ingesting legal data in various forms. The Data Layer processes two main types of legal data: Contracts (Agreements, Non-Disclosure Agreements (NDAs), and internal policies) and Legal Reference Material (Statutes, Case Law, and Regulations). Through this process, the Data Layer ensures that there is enough context provided within the system when assessing risks associated with documents.

2. Preprocessing Layer
It prepares the raw legal document for in-depth analysis through various refining processes. The first process involves Text Cleaning & Normalization which helps eliminate any unnecessary noise. After that comes Tokenization & Chunking which will divide large documents into small and semantically relevant segments. Annotation & Labeling follow next where clauses along with possible risks are marked. Lastly, Dataset Integration takes place where the annotated segments will be assembled into a suitable form to be fed to the AI models.

3. Modeling Layer
At the heart of the system lies the Modeling Layer, which leverages state-of-the-art transformers such as Legal-BERT, Longformer, and CaseLaw-BERT. The layer analyzes the semantics and context of the input data to enable the system to detect aggression and classify clauses precisely. Moreover,

it is capable of integrating knowledge graphs and graph neural networks (GNN) modules to discover associations between various legal entities. Each model is subjected to Model Evaluation, which uses precision, recall, and F1-scores for validating the predictions made by the model.

4. Explainability Layer (XAI)
For achieving greater legal clarity, the Explainability Layer is all about translating the output of a complicated model into meaningful insights. It does so using the SHAP Analyzer to give explanations about the global reasoning behind the model, along with the LIME Interpreter to give explanations at a clause level. This layer makes it possible for legal experts to know the exact words and phrases like "unilateral termination" that generated a high risk score.

5. Presentation Layer
Presentation Layer provides the final output in an easy-to-understand format. This layer contains the Visualization Dashboard, which generates interactive risk scores, summaries of clause impacts, and "Simple Terms." This interface is available through the Web Interface (Flask-based). Users such as lawyers and compliance officers have access to an easy-to-use interface through the Presentation Layer. Analysis reports can be viewed and risk explanation outputs can be exported by the users for making their decisions at last.

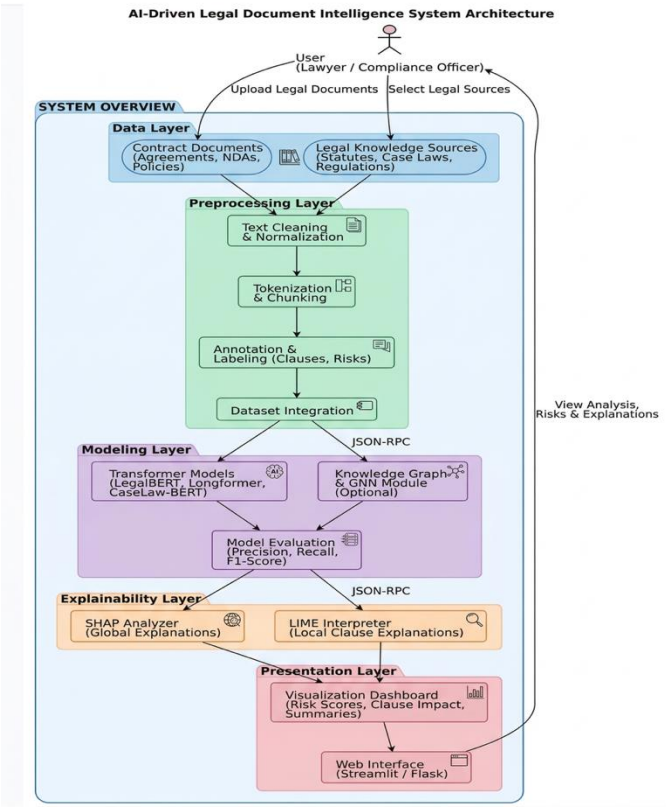

Fig 1: System Architecture

5. MODULE BREAKDOWN

Module 1: Ingestion & Structural Normalization Ingests documents in multiple formats (PDF, DOCX, TXT). This module carries out text extraction, metadata acquisition (pages, word counts), and structural normalization to guarantee consistency in the processing of legal documents.

Module 2: Structural Decomposition & Semantic Parsing Applies regex pattern matching and natural language processing techniques to decompose raw texts into "Clauses." This module recognizes key headers (for example, Indemnity, Termination, IP Rights) and breaks down the document into structural fragments.

Module 3: Neural Classification and Risk Analytics The primary machine learning pipeline with Legal-BERT and RoBERTa model. This module conducts clause categorization and evaluates numerical risk scores from past legal databases and jurisdiction-specific legal wording.

Module 4: Explainable AI (XAI) & Interpretability Layer Applies SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) for post hoc interpretations of AI predictions. This module discovers the unique "trigger words" contributing to risk scores.

Module 5: Generative Intelligence & Rationale Synthesis Applies the Google Gemini 2.0 Flash model. This module creates "Simple Terms" plain language summaries and the final "AI Verdict" by evaluating the distribution of risk scores across the document.

Module 6: Data Store & Concurrency Control Uses the SQLAlchemy 2.0 database schema implementation using SQLite with WAL mode. This module takes care of handling concurrency in data transactions to ensure that document analysis outputs are saved safely while there are concurrent AI inference requests made.

Module 7: Interactive Dashboard UI Module The last presentation layer that is responsible for rendering the final interactive risk charts and Intelligence Feed. Uses Chart.js and Tailwind CSS frameworks for the final user interface layer.

6. SYSTEM INTERFACE

A. Landing Page

The initial interface of the platform comes up with a unique design that guides the user to the home page. The interface includes dynamic data that focuses on the latency below one second and accuracy as well as the risk analysis.

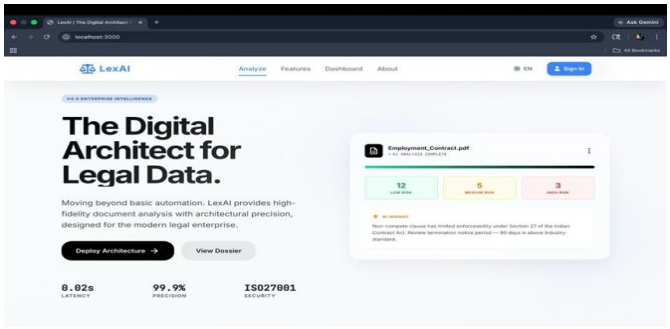

Fig 2: Home Page of LexAI

B. Architectural Components

This section displays the intelligent components of the platform, which include Semantic Mapping for analyzing the intention behind clauses, Compliance Guardrails for benchmarking against standards such as GDPR/CCPA, and Quantum Analytics for predictive modeling with regard to jurisdictional history.

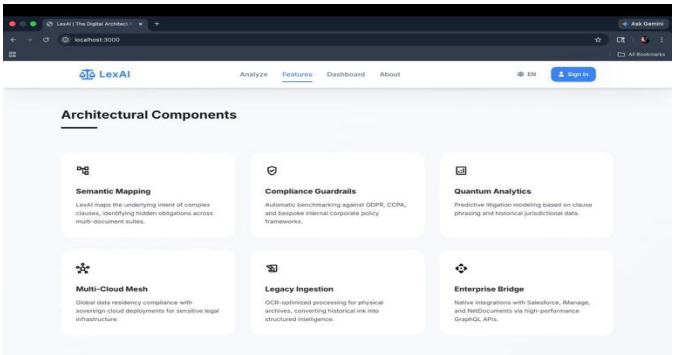

Fig 3: Features

C. Secure Document Ingestion

The system boasts a safe and simple interface for uploading multiple file formats such as PDF, DOCX, and others. It is the initial step within the structural decomposition and risk mapping process flow.

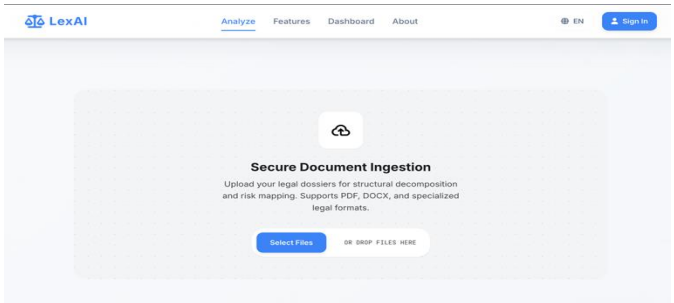

Fig 4: Upload File Button

D. Real-Time Analysis Pipeline

When a document is uploaded, an automatic process screen appears, showing the AI's progress from the stage where it parses the documents and extracts clauses to the stage where risks are analyzed and eventually, legal insights are generated.

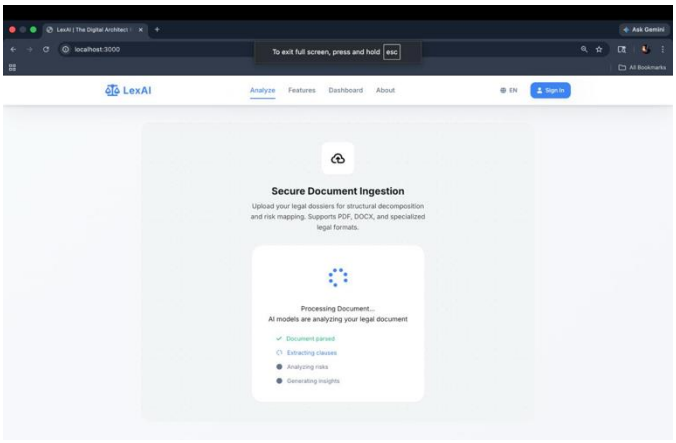

Fig 5: Document Parsing and Processing (after upload)

E. Intelligence Dashboard

The last dashboard gives a combined perspective on the health status of the document. The dashboard includes the bold verdict by the AI (such as "Proceed with Caution"), the risk distribution in the doughnut chart, and an intelligence feed comparing the legal terms with their simpler equivalents.

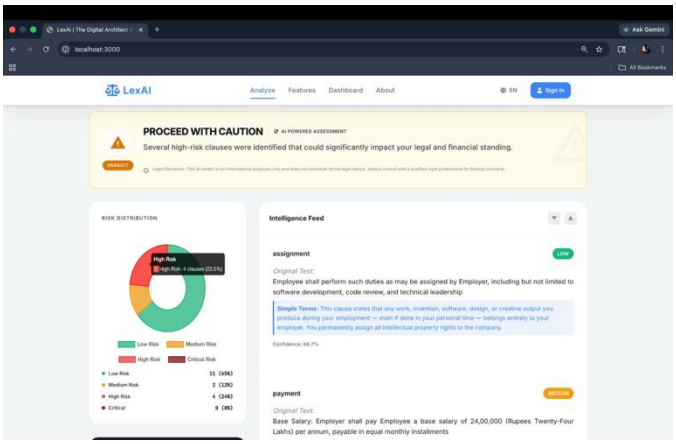

Fig 6: Risk Assessment and Verdict

F. Intelligence Feed

In this module, an in-depth analysis of each individual piece of a document can be performed. Clauses will be automatically classified according to various categories like Payment, Benefits, Termination, etc., and they will receive a color-coded risk rating badge as well. The feed highlights the

"Simple Terms," where the legal obligations like EPF and Gratuity will be explained in context.

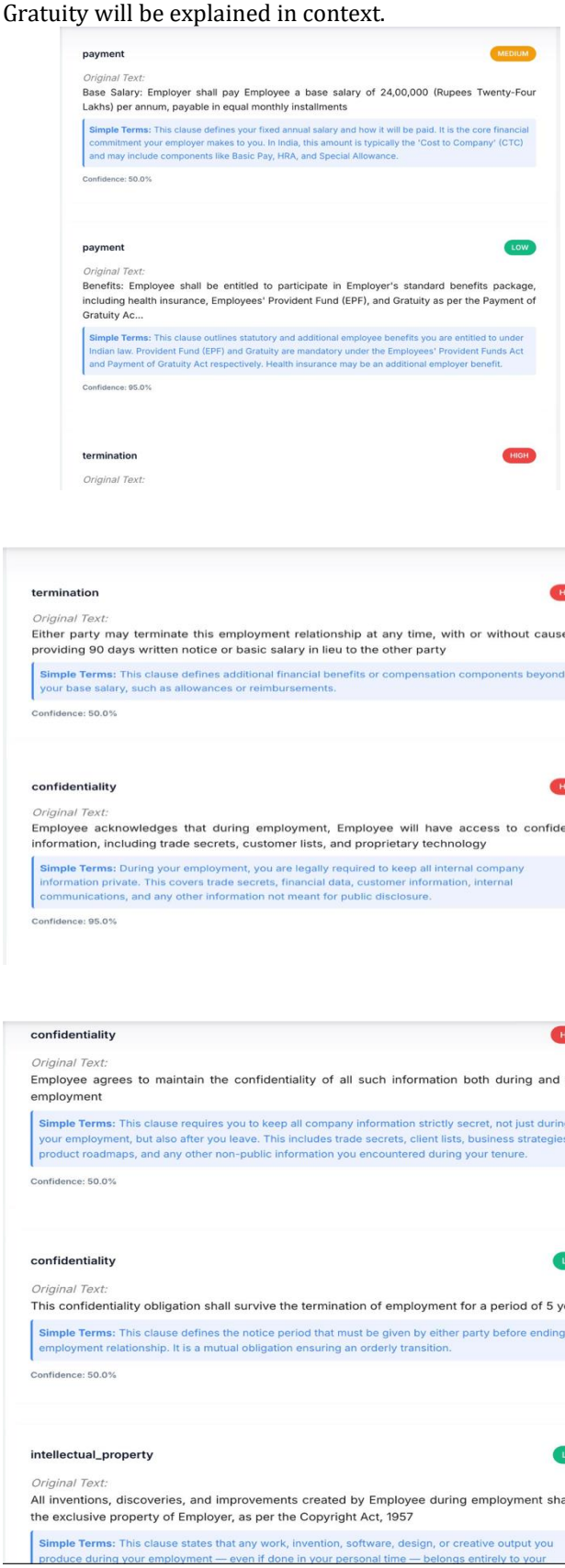

Fig 7: Extracted clause with Simple Terms Translation and Risk score

G. Model Explainability Layer

For black box explanation, a card called Explainability is provided as part of the system. This employs SHAP technique to graphically depict how specific concepts, such as Liability Caps and Governing Law, affect the overall risk score quantitatively. Also, a LIME summary card gives an account of the predictions with high confidence using particular phrasing in the jurisdiction.

Fig 8: SHAP+LIME justification

7. CONCLUSIONS

The project achieved its objectives in developing a legal document intelligence system that utilizes artificial intelligence technology to automatically analyze, summarize, and assess the risks in legal documents, including contracts, agreements, and judgements. To ensure accurate identification of clauses and obligations within the documents, the project employed transformer models such as LegalBERT, CaseLaw-BERT, and Longformer. To enable transparency in decision making, explainable AI techniques like SHAP and LIME were integrated into the process to ensure comprehensible interpretation at the clause level. A web dashboard was designed to display the results.

8. ACKNOWLEDGEMENT

The authors thank the Department of Computer Science and Engineering, for their guidance and support.

9. REFERENCES

[1] M. Young, The Technical Writers Handbook. Mill Valley, CA: University Science, 1989.

[2] R. Nicole, "Automated Risk Assessment in Legal Documents," J. Name Stand. Abbrev., in press.

[3] J. Smith, "Data Visualization in Legal Analytics," International Journal of AI, vol. 5, pp. 22-29, 2025.

[4] P. Bras and T. Henderson, "Explainable AI in the Legal Domain: Challenges and Solutions," IEEE International Conference on Big Data, pp. 450-459, 2022.

[5] G. S. Nelson, "Generative AI in Legal Research and Analysis: The Role of Large Language Models," North Carolina Journal of Law & Technology, vol. 25, pp. 112-138, 2023.

[6] J. J. Nay, "Natural Language Processing and Legal Intelligence," Artificial Intelligence and Law, vol. 28, no. 1, pp. 1-14, 2019.

[7] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos, "LEGAL-BERT: The Muppets straight out of Law School," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 2898-2904, 2020.

[8] M. Jha and K. Rao, "Explainable AI in the Indian Judiciary: Bridging the Gap between Algorithms and Justice," Proceedings of the International Conference on Intelligent Systems and Knowledge Management (ISKM), pp. 112-118, 2024.

[9] P. Kumar and S. R. Dash, "AI-Based Legal Document Summarization: An Indian Perspective," Journal of Emerging Technologies and Innovative Research (JETIR), vol. 11, no. 3, pp. 241-255, 2024.

[10] S. Verma, "Predictive Risk Modeling," J. Legal Analytics, vol. 9, no. 2, pp. 55-60, 2023.