International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4289 - 4294
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0558

AI Based Judicial Data Mining System

Harshad R. Chavan[1], Ganesh A. Bedake[2], Atharv D. Yadav[3], Tanmay M. Kalbhor[4], Nimish K. Sawant[5], Asma N. Mulla[6]

[1,2,3,4,5]UG Scholar, Department of CSE, Yashoda Technical Campus, Satara, Maharashtra, India.

[6]Assistant Professor, Department of CSE, Yashoda Technical Campus, Satara, Maharashtra, India.

Emails: *harshadchavan4788@gmail.com[1], ganeshbedake7285@gmail.com[2], atharvyadav1135@gmail.com[3], tanmaykalbhor21@gmail.com[4], nimishsawant153@gmail.com[5], asmamulla_engg@yes,edu.in[6]*

Abstract

The rapid expansion of digital legal information in India has created a strong demand for intelligent systems capable of supporting structured judicial research, legal analytics, and data-driven decision-making. In addition, the complexity of legal case documents makes information retrieval and judicial analysis challenging for researchers and legal professionals. In legal system design, the capability to efficiently process and analyze judicial data plays a vital role. Today, intelligent analytical systems are progressively being introduced at earlier stages of legal research with Artificial Intelligence and Natural Language Processing techniques. In the current research, an AI-Based Judicial Data Mining and Analytics System is developed. This system is implemented using PyQt6 and SQLite for enhancing accessibility, transparency, and analytical depth within the judicial domain. The proposed system integrates Natural Language Processing (NLP) and machine learning techniques for semantic case search using sentence embeddings and FAISS-based similarity indexing. It processes judicial case documents and extracts structured information such as parties, judges, court details, and case outcomes. Judicial analytics is the significant part of improving legal research and decision support systems. Finding important citation relationships in legal documents is utilized as a driving force to develop a superior analytical platform. Judge performance analytics, citation network analysis using PageRank methods, and outcome prediction modules have to be explored to comprehend the impacts of intelligent legal analytics and data-driven judicial insights.

Keywords: *Citation analysis; Judicial data mining; Legal analytics; Machine learning; Natural language processing.*

1. Introduction

Judicial data in India is increasing rapidly in both volume and complexity, with millions of pending and resolved cases distributed across multiple levels of courts. Despite the availability of digital records, a significant portion of legal data remains unstructured, making it difficult to navigate, interpret, and analyze efficiently. Traditional keyword-based search systems are often insufficient, as they fail to capture semantic relationships within legal text, leading to time-consuming and less effective legal research processes. As legal professionals, researchers, and policymakers increasingly rely on data-driven insights, the need for intelligent and automated judicial information systems has become critical. This research proposes an AI-Based Judicial Data Mining and Analytics System, a standalone desktop platform designed to systematically organize, analyze, and visualize judicial data. Unlike conventional web-based systems, the application supports complete offline functionality, ensuring privacy, portability, and uninterrupted access to legal information. Developed using PyQt6 for the graphical user interface and SQLite for efficient local data storage, the system provides a lightweight yet

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4289 - 4294
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0558

powerful environment for legal data exploration. The system integrates advanced Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques to enhance legal research capabilities. It employs semantic search using sentence embeddings and FAISS-based similarity indexing to retrieve contextually relevant cases. Additionally, it incorporates judge performance analytics, citation network analysis using graph-based methods, and machine learning models for predicting case outcomes. An AI-powered assistant further improves usability by generating simplified explanations of complex legal cases. The primary objective of this project is to bridge the gap between raw judicial datasets and actionable legal insights by combining structured data modeling, intelligent search mechanisms, and analytical tools. By doing so, the system aims to improve efficiency, accessibility, and decision-making within the legal domain, while also providing a scalable foundation for future AI-driven legal informatics applications.

2. Literature Review

The Judicial analytics and legal information systems have been widely developed across the globe to support legal research and decision-making. Prominent platforms such as LexisNexis and Westlaw provide comprehensive access to legal databases, case laws, and analytical tools. In India, initiatives such as the eCourts project and the National Judicial Data Grid (NJDG) aim to digitize court records and improve accessibility. However, these systems are often subscription-based, dependent on continuous internet connectivity, and offer limited customization for individual users. Recent advancements in legal informatics research have focused on three major areas. The first area involves data extraction and structuring, where techniques are developed to convert unstructured judicial documents into structured formats for efficient querying and analysis. The second area focuses on legal analytics and visualization, enabling users to identify patterns, trends, and relationships within case data through statistical and graphical methods. The third area explores the use of Artificial Intelligence and machine learning for predictive modeling, including case outcome prediction, legal document classification, and case similarity detection. More recently, Natural Language Processing (NLP) techniques such as transformer-based models and sentence embeddings have been applied to improve semantic understanding of legal text. These approaches allow systems to move beyond keyword-based search and enable context-aware retrieval of relevant legal information. Additionally, graph-based methods such as citation network analysis using PageRank have been employed to identify influential precedents and understand relationships between cases. Despite these advancements, most existing systems are designed for large-scale institutional use and lack affordability, offline accessibility, and flexibility for independent users such as students, advocates, and researchers. There remains a significant gap in the availability of lightweight, standalone systems that integrate semantic search, analytics, and AI capabilities in a unified platform. The proposed system addresses this gap by providing an offline, customizable judicial analytics platform that combines NLP-based semantic search, machine learning-driven prediction, and graph-based citation analysis, thereby offering an accessible and extensible solution for modern legal research.

3. System Architecture and Methodology

The Methods section describes the design and components of the AI-Based Judicial Data Mining and Analytics System. Each module below presents sufficient technical detail to allow replication by a qualified reader. Previously established procedures are cited where applicable [1–12].

3.1. Technology Stack

The proposed system is developed using a combination of modern software and AI technologies: Frontend GUI using PyQt6 (Qt for Python) for building an interactive desktop interface; Database using SQLite, a lightweight file-based database for offline storage and fast querying; NLP Model using Sentence Transformers (MiniLM) for generating semantic embeddings; Vector Search Engine using FAISS for efficient similarity search; Machine Learning using Scikit-learn-based models for outcome prediction; Graph Analysis using NetworkX for citation network modeling; Visualization using Matplotlib for charts and analytics; and a Dataset of judicial case documents

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4289 - 4294
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0558

extracted from PDFs and structured into a relational database.

3.2. Data Processing and Dataset Structure

The system processes judicial case documents through multiple stages including extraction, cleaning, and structuring. The dataset captures a multidimensional representation of legal cases, including: case metadata (case number, year, court type, and judgment date); party information (petitioner and respondent); judicial attributes (judge names, bench details, and court); case outcome (allowed, dismissed, or partially allowed); textual data (full judgment text and generated summaries); and additional attributes such as extracted citations and clustering labels. This structured representation enables efficient indexing, querying, and analytical processing of legal data [13-15].

3.3. Data Processing Pipeline

The system follows a structured pipeline for transforming raw judicial documents into analyzable data. (1) Data Extraction: case PDFs are processed using text extraction tools to obtain raw textual content. (2) Data Cleaning: extracted text is normalized by removing noise, formatting inconsistencies, and irrelevant tokens. (3) Information Extraction: key entities such as parties, judges, court names, and outcomes are extracted using rule-based and pattern-matching techniques. (4) Data Storage: the processed data is stored in relational tables within SQLite for efficient access and querying.

3.4. Semantic Search System

Traditional keyword-based search methods are insufficient for legal text due to its complexity and variability. To address this, the system employs semantic search using Natural Language Processing. Case documents are converted into dense vector representations using Sentence Transformers. These embeddings are indexed using FAISS, enabling efficient similarity-based retrieval. User queries are also converted into embeddings and matched against stored vectors to retrieve contextually relevant cases. This approach significantly improves search accuracy by capturing semantic relationships rather than relying on exact keyword matches.

3.5. Judge Analytics Module

The system analyzes judicial performance by extracting judge names from case data and computing statistical metrics such as total number of cases handled, distribution of outcomes (allowed, dismissed, partially allowed), and decision patterns and success rates. Judges are ranked using weighted scoring methods that combine case volume and outcome performance. This module helps identify specialization trends and decision behaviors.

3.6. Citation Network Analysis

To identify influential legal precedents, the system constructs a citation network from case texts. Citations are extracted using pattern matching and used to build a directed graph where nodes represent cases or citations and edges represent references between cases. NetworkX is used to compute graph-based metrics such as in-degree and PageRank. These metrics help determine the influence and importance of specific cases within the legal network.

3.7. Outcome Prediction Model

A machine learning-based prediction module is implemented to estimate case outcomes. The model analyzes textual features of case descriptions and generates probabilities for outcomes such as Allowed, Dismissed, and Partially Allowed. The system also provides confidence scores, enabling users to understand the reliability of predictions.

3.8. Case Visualization and Analytics

The system provides detailed case visualization, displaying structured information such as parties, judges, court, summary, and outcome. Analytical insights are generated using visualization tools, including outcome distribution charts, case trends over time, and statistical summaries of judicial data. These visualizations help users identify patterns and trends within legal datasets.

3.9. AI Legal Assistant

An AI-powered assistant is integrated to enhance usability. It generates simplified explanations of legal cases, enabling users to understand complex judicial content more easily. This feature improves accessibility for non-experts and students.

3.10. Reporting and Exporting

The system supports exporting case details into structured formats such as PDF reports. This functionality is useful for academic work, legal documentation, and presentations Shown in Table 1.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4289 - 4294
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0558

Table 1 Technologies and Functional Modules of the AI-Based Judicial Data Mining and Analytics System

Module	Technology	Purpose
User Interface	PyQt6	Provides an interactive desktop-based judicial analytics platform
Database Management	SQLite	Stores judicial case records
Case Document Processing	NLP	Extracts and preprocesses legal text data
Semantic Case Search	Sentence Embeddings + FAISS	Enables intelligent similarity-based case retrieval
Data Extraction Module	Entity Recognition Techniques	Extracts parties, judges, court names, and outcomes
Citation Analysis	Graph-Based Analysis (PageRank)	Identifies influential legal cases and citation relationships
Judge Performance Analytics	Statistical & Analytical Methods	Evaluates judicial decision patterns and performance

4. Results and Discussion

4.1. Results

The proposed AI-Based Judicial Data Mining and Analytics System was evaluated based on performance, usability, and analytical capabilities. The system demonstrated efficient handling of large judicial datasets, with an average semantic search response time of less than 0.3 seconds using FAISS-based similarity indexing. Compared to traditional keyword-based approaches, the semantic search module provided significantly more relevant and context-aware results. The data ingestion and processing pipeline successfully extracted and structured key information such as parties, judges, court details, and case outcomes from judicial documents. The generated summaries enabled quick understanding of case content, improving user efficiency during legal research. The Judge Analytics module produced meaningful insights into judicial performance, including case distribution, outcome ratios, and decision patterns. This helped identify trends in judicial behavior and specialization. Similarly, the Citation Network Analysis module effectively identified influential cases using graph-based metrics such as in-degree and PageRank, highlighting key precedents within the dataset.

4.2. Discussion

The visualization components, including charts and dashboards, helped users analyze judicial trends and case patterns effectively. User feedback indicated that the interface was intuitive and visually appealing. The modular architecture also supports future integration of advanced AI features and real-time data sources. Overall, the system demonstrates strong usability and significant potential for improving legal research and analytics Shown in Figure 1.

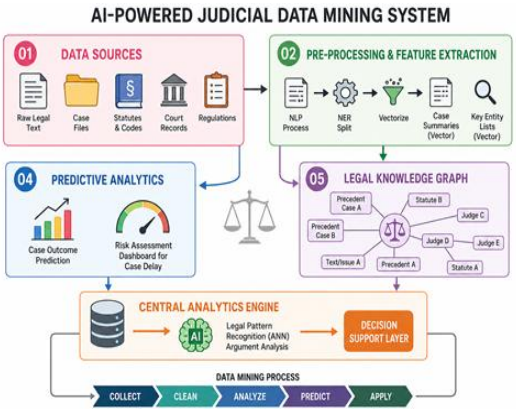

Figure 1 Architecture of an AI-Powered Judicial Data Mining System

Conclusion

The rapid growth of digital judicial records and legal documentation has made traditional methods of legal analysis increasingly inadequate for achieving timely, accurate, and efficient decision-making. The AI-Based Judicial Data Mining and Analytics System presented in this study demonstrates how modern Artificial Intelligence (AI), Natural Language Processing (NLP), and machine learning techniques can significantly enhance the process of legal research and analysis. The system integrates semantic search using embeddings and FAISS, enabling

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4289 - 4294
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0558

context-aware retrieval of relevant cases beyond conventional keyword-based methods. In addition, modules such as judge analytics, citation network analysis, and outcome prediction provide deeper insights into judicial patterns, case relationships, and decision trends. The inclusion of an AI-powered assistant further improves accessibility by simplifying complex legal information for users. By automating data extraction, improving information retrieval, and providing analytical insights, the system has the potential to reduce research time, improve decision-making, and enhance transparency within the judicial ecosystem. Furthermore, its modular and scalable architecture allows for future integration with real-time judicial databases and advanced AI models. Overall, this work highlights the practical applicability of AI-driven approaches in the legal domain and establishes a strong foundation for the development of intelligent, data-driven judicial systems that can support judges, lawyers, researchers, and policymakers in the evolving landscape of legal informatics.

Acknowledgements

The authors would like to express their sincere gratitude to the faculty members of the Department of Computer Science and Engineering for their continuous guidance, valuable suggestions, and encouragement throughout the development of this project. Their support played a vital role in shaping both the technical and research aspects of this work. The authors also extend their heartfelt appreciation to their project guide for providing constant mentorship, insightful feedback, and direction during the course of this research. Special thanks are extended to friends and peers for their cooperation, constructive feedback, and assistance during the testing and evaluation phases of the system. Their contributions helped improve the quality and usability of the final implementation. Finally, the authors express their deep gratitude to their family members for their unwavering support, patience, and motivation throughout the entire research and development process.

References

[1]. Government of India, "eCourts Services: National Judicial Data Grid (NJDG)," Ministry of Law and Justice. [Online]. Available: https://ecourts.gov.in
[2]. M. Medvedeva, M. Vols, and M. Wieling, "Using machine learning to predict decisions of the European Court of Human Rights," Artificial Intelligence and Law, vol. 28, pp. 237–266, 2020.
[3]. S. K. Nigam et al., "NyayaAnumana & INLegalLlama: The Largest Indian Legal Judgment Prediction Dataset and Specialized Language Model," arXiv preprint, 2024.
[4]. "Artificial Intelligence in Legal Systems," International Journal of Law, Policy and Social Review, vol. 6, no. 3, pp. 164–168, 2024.
[5]. R. Sharma, Justice through Technology: An Indian Approach, 2nd ed. New Delhi, India: LexisNexis, 2019.
[6]. K. Joshi, Technology and Law: Challenges for the Indian Judiciary. Lucknow, India: Eastern Book Company, 2022.
[7]. D. Chalkidis, I. Androutsopoulos, and N. Aletras, "Neural Legal Judgment Prediction in English," Proceedings of ACL, 2019.
[8]. N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos, "Predicting judicial decisions of the European Court of Human Rights," PeerJ Computer Science, 2016.
[9]. S. Katz, M. Bommarito, and J. Blackman, "A general approach for predicting the behavior of the Supreme Court of the United States," PLOS ONE, 2017.
[10]. Ministry of Electronics and Information Technology (MeitY), "National Strategy for Artificial Intelligence," Government of India, 2018.
[11]. NITI Aayog, "Responsible AI for All: National Strategy for Artificial Intelligence," Government of India, 2021.
[12]. T. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," arXiv, 2013.
[13]. Supreme Court of India, "SUPACE: Supreme Court Portal for Assistance in Court Efficiency," Supreme Court of India, 2021.
[14]. NITI Aayog, "National Strategy for Artificial Intelligence #AIforAll," Government of India, New Delhi, 2018.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4289 - 4294
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0558

[15]. Ministry of Electronics and Information Technology (MeitY), “Digital Personal Data Protection Act and AI Governance Initiatives in India,” Government of India, 2023.