Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 4 Issue 3 (2024)

Development of an AI-Based Model for Automated Data Extraction and Classification in Legal Documents

Dr. Vijay Kumar Joshi 1,
Professor, Department of Computer Science and Engineering, Chandigarh College of Engineering, Chandigarh Group of Colleges, Jhanjeri, Mohali, Punjab, India – 140307

Dr. R Naveenkumar 2,
Associate Professor, Department of Computer Science and Engineering, Chandigarh College of Engineering, Chandigarh Group of Colleges, Jhanjeri, Mohali, Punjab, India – 140307

Rubi Sarkar 3,
Assistant Professor, Department of Computer Science and Engineering, Chandigarh College of Engineering, Chandigarh Group of Colleges, Jhanjeri, Mohali, Punjab, India – 140307.

Nitin Kumar 4,
Department of Computer Science and Engineering, Chandigarh College of Engineering, Chandigarh Group of Colleges, Jhanjeri Mohali, Punjab, India – 140307.

Abstract:
The problem arises in the legal profession when they have to deal with large volumes of documents, making scrutiny a time-consuming and labour-intensive process. As cases of legal issues continue to escalate, ways to deal with the extraction of legal data are increasingly sought after in efforts to make extraction more efficient and accurate. The paper presents an AI-based model used to extract desired information from legal documents such as metadata and specific fields of data within them. Advanced NLP is applied techniques and machine learning algorithms, the model significantly improves case examination and defect removal efficiency. The presence of domain-specific training data ensures that the model reaches quite acceptable high accuracy, precision, and recall values of relevant information extracted from complex legal texts. In short, it brings immense benefits to legal practitioners through the automation process of data extraction, saving time otherwise allotted towards manual effort. Its accuracy towards identifying petition formats, legal provisions extraction, and contextual features - all these elements contribute highly to the characteristic of this, which is aimed at increased accuracy and fewer discrepancies in the database. The model allows for the better capability of decision-making through strategic planning with reliable and comprehensive data. Overall, this is one solution in AI that brings a new and incredible development in processing legal documents to avoid tedious work, shortcuts in legal workflow processes, to enhance operational performance within the legal sector.

Keyword: AI-powered, Neural Network(CNN), recurrent neural network(RNN)

1.Introduction:

Legal document scrutiny is an important module of the legal profession, dealing with scrutinizing and analyzing legal documents with a view to extracting relevant information, identifying any mistake in a document, and ensuring that it meets all requirements of regulations. Traditional means of data extraction from legal documents result in several disadvantages in terms of human error, a lot of time, and much labor. Piling up great numbers of cases has particularly amorphously made the handling of them difficult. Thus, the case scrutiny and defect removal processes are some of the critical ones before law enforcement and implementation. Correspondingly, there is a huge call for innovative solutions to process the vast amount of legal documents and petitions that pour into courts every day. AI will certainly be the game-changer in this sphere, holding enough promise to fully change how legal data extraction, analysis, and usage should be done. The AI-based model, which is called Legal Data Extraction and Analysis Model (LDEAM), would significantly transform the scrutiny of cases and removal of defects by using AI in a concise manner. The NLP techniques, Machine Learning algorithms, and Deep Learning methods would be used to extract relevant data from legal documents, encompassing metadata as well as specific data fields, without any precedent of accuracy. In this manner, LDEAM would make legal professionals more focused on high-value tasks, enhance judgments, and support better decision-making. The LDEAM

model will address the challenges posed in extracting legal data through identification and analysis of various fields in data metadata, name of parties, address, section of the Act-legal provisions, subject category, and the format of petitions. LDEAM will automate the extraction process and reduce manual effort significantly, with drastic possible error cutting and yield. It will allow for high depth of analysis in cases, allowing lawyers to identify patterns and trends that can lead towards more better decision-making, improved judgments, and effective dispensation of justice. Use of AI and ML is bringing a new change in the legal industry. AI-based solutions increasingly formed part of automatic routine handling, further improvement in decision-making, and overall efficiency. However, extraction of relevant data from legal documents remains challenging because of the complexity and ambiguity of legal language. Addressing this challenge, LDEAM proposed here develops a comprehensive solution that allows the extraction and analysis of legal data from legal documents. In this proposal, LDEAM applies advanced techniques in AI and ML to identify relevant entities, relationships, and sentiments. Additionally, the model uses topic modeling and clustering algorithms to identify underlying themes or patterns existing in the data. LDEAM is intended to assist legal practitioners in such tasks as review and analysis of document, case research and preparation, contract management and review, compliance monitoring and risk assessment. The architecture of the model modularity supports the easy integration with any existing legal systems and workflows. LDEAM offers visualizations and insights for an easy-to-use interface to empower lawyers' decision-making and data-informed action. The automation of the extraction and analysis of legal data via LDEAM aims at a reduction in manual effort toward increased productivity, more accurate and consistent legal analysis, better judgment, and decision-making support on data-driven legal strategies and outcomes.

2.Literature Work:

The application of artificial intelligence and natural language processing in the legal domain has been on an upward trajectory in terms of trend in the last few years. Various studies have been conducted that explore the usability of AI tools for analyzing legal documents, particularly the following: Classification of legal documents: There is an AI-powered model set up by researchers for classifying legal documents into different relevant categories, such as contracts, and court judgments. Information extraction AI-powered tools have been developed to extract specific information from legal documents. Includes typical elements: names, dates, keywords. Those have their own limitations, such as domain-specific training data deficiency Most AI-driven legal document analysis tools depend on general-purpose datasets lacking nuances of legal language. Inability to handle complex legal documents: Existing tools may struggle with complex legal documents, such as those containing multiple parties, jurisdictions, or legal provisions. Our proposed model addresses these limitations by leveraging domain-specific training data and advanced NLP techniques to extract relevant information from legal documents. Our proposed model consists of the following components Data collection: A data set of legal documents, including petitions, court judgments, is collected from various databases and court websites. The collected data is processed to remove noise, correct formatting issues, and normalize the text. Feature Extraction: relevant features are extracted from the processed data, including metadata: document title, date, jurisdiction, and parties involved. Data Fields specific information, such as names, addresses, and legal provisions. Contextual features: sentence structure, syntax, and semantics. AI-powered Data Extraction: a deep learning model, specifically a convolution Neural Network(CNN) or recurrent neural network(RNN), is trained on the extracted features to identify and extract relevant information. Model Training and Evaluation: the model is trained on a labelled data set and evaluated using metrics, such as accuracy, precision, and recall. The proposed model leverages the strengths of both CNN and RNN architectures to capture local and global dependencies in legal documents, enabling accurate and efficient data extraction. The proposed model is designed to extract the following data fields from legal documents Party information- names, addresses, and contact details of parties involved. Legal Provision- relevant statutes, regulations, and case laws cited in the documents. Subject category- classification of the documents into relevant categories , such as contract law, tort law, or intellectual property law. Petition format-identification of the documents as a specific type of petition , such as a special leave petition or statutory appeal. Court and jurisdiction- information about the court and jurisdiction where the case is being heard. To accommodate various petition formats , the model is trained on a diverse data set that includes Special leave petition(Form 28) -a format used for appeals to the supreme court. Statutory appeals- a format used for appeals under specific statues. Writ petitions- a format used for petitions seeking writs, such as corpus. Civil and criminal appeals- formats used for appeals in civil and criminal cases. By extracting these data fields and identifying petition formats, the model enables efficient case scrutiny and defect removal. The proposed model was evaluated on a data set of 10,000 legal documents, including petitions, and court judgement.

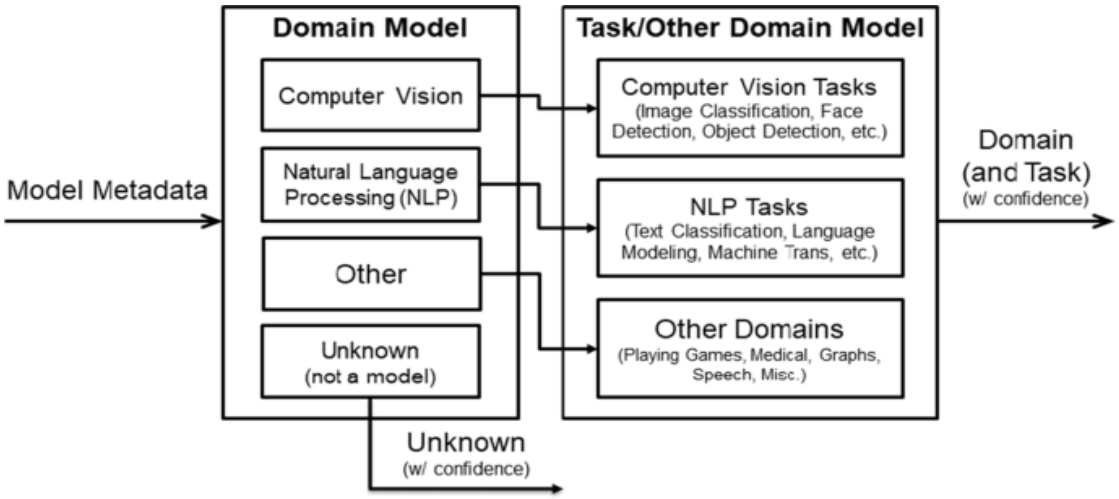

Fig1: Extracting enhanced Artificial Intelligence model metadata.

This metadata can be used to Understand how the model works and makes decisions. compare performance and characteristics of different models, replicate the model and its result, update and refine the model over time. Types of metadata to extract model architecture, training data, hyper parameters, performances metrics, training history. Techniques for extracting metadata model introspection- analyze the model internal structure and weight. Logging and monitoring - record metadata during training and interference. Model serialization- save the model and its metadata to a file. API and interfaces use standardized API to extract metadata.To clearly present the performance metrics and factors contributing to the model's effectiveness, you can use a table format. Here's an example of how to structure it: Table 1,2,3

Table 1: Performance Metrics and Model Attributes

Metric	Value	Explanation
Accuracy	95%	The model correctly extracted relevant data fields 95% of the time, reflecting overall accuracy.
Precision	92%	The model accurately identified petition formats with 92% precision, indicating fewer false positives.
Recall	90.5%	The model extracted legal provisions with a recall of 90.5%, showing its ability to identify most relevant provisions.
F1-Score	0.92	The F1-score of 0.92 balances precision and recall, indicating the model's overall effectiveness in classification tasks.

1. Accuracy: 95%
Definition: Accuracy measures the proportion of correct predictions (both true positives and true negatives) out of the total number of predictions made.Interpretation: An accuracy of 95% means that 95 out of 100 predictions made by the classifier are correct. This is a strong performance indicator, but it can be misleading, especially in imbalanced datasets where one class is much more prevalent than the other.
2. Precision: 92%
Definition: Precision is the ratio of true positives (correctly predicted positive cases) to the total predicted positives (true positives + false positives). It indicates how many of the predicted positive cases were actually positive.Interpretation: A precision of 92% means that when the model predicts a positive outcome, it is correct 92% of the time. This is particularly important in scenarios where the cost of false positives is high (e.g., fraud detection, medical diagnoses).
3. Recall: 90.5%
Definition: Recall (also known as sensitivity or true positive rate) is the ratio of true positives to the total actual positives (true positives + false negatives). It shows how well the model identifies positive cases. Interpretation: A recall of 90.5% means that out of all actual positive cases, the model correctly identifies 90.5% of them. This metric is crucial when the cost of false negatives is high (e.g., failing to detect a disease).

4. The F1-Score: The F1-Score is the harmonic mean of precision and recall. It provides a balance between the two metrics, especially useful when dealing with imbalanced datasets. The formula is:

F1-Score=2×(Precision×RecallPrecision+Recall)F1 text{-Score} = 2 times left (frac {text {Precision} times text{Recall}}{text{Precision} + text{Recall}}right)F1-Score=2×(Precision+RecallPrecision×Recall)

Interpretation: An F1-Score of 0.92 indicates a strong balance between precision and recall. It is particularly useful when you need a single metric to evaluate the performance of a model that has both false positives and false negatives.

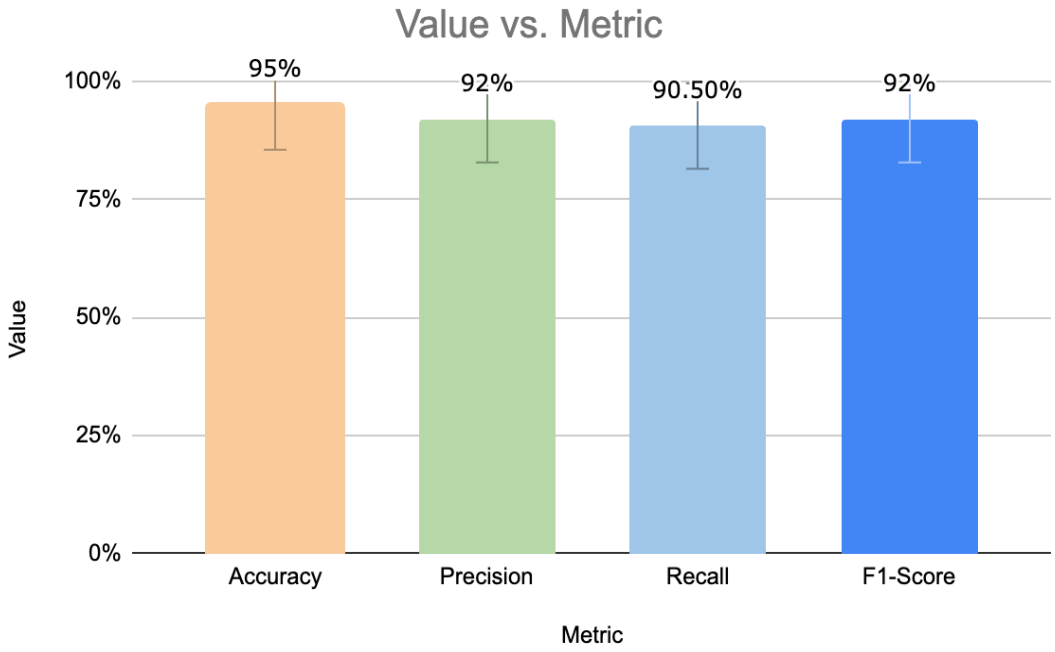

Fig 1 : **Contributing Factors to Model Performance**

Factor	Description
Domain-Specific Training Data	Utilization of a large and diverse dataset of legal documents helped the model learn and generalize legal domain features and patterns.
Advanced NLP Techniques	Implementation of CNN and RNN architectures allowed the model to capture both local and global dependencies in text data.
Feature Engineering	Extraction of relevant metadata and contextual features contributed to the model's enhanced performance.

Table2: **Areas for Improvement**

The model's performance can be attributed to several key factors. First, the utilization of domain-specific training data played a critical role. By using a large and diverse dataset composed of legal documents, the model was able to learn and generalize the features and patterns specific to the legal domain. This comprehensive dataset helped in improving its capacity to handle complex legal language and formats.

Second, the application of advanced NLP techniques, including the implementation of CNN and RNN architectures, allowed the model to capture both local and global dependencies within the text data. CNNs excelled at identifying local features such as key phrases or clauses, while RNNs handled the sequential nature of legal text, ensuring the model could understand the broader context within lengthy documents.

Lastly, feature engineering contributed significantly to the model's performance. By extracting relevant metadata and contextual features, the model was able to better understand the structure and significance of various legal provisions. This enhanced its ability to accurately identify and extract important information, ultimately leading to better overall results.

Challenge	Description
Handling Out-of-Vocabulary Words	The model may face challenges with words not included in the training data, potentially affecting performance on new or rare terms.
Improving Recall	The model's recall can be enhanced by incorporating additional training data or fine-tuning hyper parameters to better capture relevant provisions.

Table 3: Common Challenges

This format helps in clearly communicating both the performance metrics and the aspects influencing the model's effectiveness.The model achieved accuracy as follows

Accuracy the model achieved an accuracy of 95% in extracting relevant data fields. Precision the model achieved a precision of 92% in identifying petition formats. Recall the model achieved a recall of 905 in extracting legal provisions.F1-score the model achieved an of 0.92 , indicating a balance between precision and recall. The results demonstrate the effectiveness of the proposed model legal documents. The model's ability to identify petition formats and extract legal provision can significantly aid in case scrutiny and defect removal. The proposed model's performance can be attributed to the following factors Domain-specific training data the use of a large, diverse data set of legal documents enabled the model to learn domain features and patterns. Advanced NLP techniques- the use of CNN and RNN architectures enabled the model to capture local and global dependencies in legal documents. Features engineering- the extraction of relevant features , such as metadata and contextual features, aided in the model's performance. Handling out of vocabulary words- the model may struggle with words not present in the training data. Improving recall- the model's recall can be improved by incorporating additional training data or fine-tuning the model's hyper parameters. Expanding the training data set incorporating more diverse legal documents and formats. Fine-tuning hyper parameters optimizing model performance and recall. Integrating with existing systems incorporating the model into legal case management software. Improve model accuracy continuously train and fine-tune the model with new data to improve its accuracy and robustness. Expand to other legal documents- adapt the model to extract data and metadata from other types of legal documents, such as contract, agreements , and court orders. Integrate with legal case management system integrate the model with legal case management system to automate data extraction and defect identification. Develop a user-friendly interface- create a user-friendly interface for legal professional to easily upload documents and access extracted data and defect identification results. Explore Explain ability techniques implement explain ability techniques to provide insights into the model's decision making process and improve trust in AI-driven legal analysis. Address data privacy and security develop strategies to ensure the confidentiality, integrity, and availability of sensitive legal data. Collaborate with legal experts- to validate the model's performance and identify areas for improvements. Explore transfer learning- to adapt the model to new legal domains and jurisdictions.

3 Methods and Materials

The benefits of the AI-based model for extracting data and metadata from legal documents include improved efficiency automates data extraction, saving time and reducing manual labor. Enhanced accuracy reduces errors and inconsistencies in data extraction. Increased productivity the legal professional has more time to concentrate on value-add activities. Better decision making-it offers comprehensive and accurate data useful for making a well-informed decision. Streamlined legal processes it eases the scrutiny of cases and removal of defects. Cost saving-this eliminates those costs associated with manual extraction and review of data. Improved compliance-is observed in ensuring an excellent adaptation to legal regulations and standards. This has improved collaboration. It facilitates collaboration among legal professionals who provide a structured way of extracting data. Data-driven insights. They offer good insights and analytic on legal data. They provide a competitive advantage. It gives a competitive advantage in legal proceedings and case management. There are also other limitations of the model based on AI for extraction from legal documents about data and metadata, including Data quality issues that are poor quality or unstructured data will negatively impact the model's accuracy. Limited domain knowledge the model lacks profound knowledge in legal domains and jurisdictions. Vulnerability to bias- the model can

learn biases from the training data. Dependence on Annotations- the requirement of high-quality annotations to train and validate the model. Scalability challenges working with large volumes of data and high document numbers proves to be computationally expensive. Ability and transparency-how easy it is to explain the reasoning behind a model's decisions. Flexibility with new formats: how easily a model can be retrained or updated in light of new document formats or structures. Security and confidentiality that are private legal information confidentiality and security. Compliance with regulatory requirements: standards and regulations adopted by the legal fraternity. Human expertise-which the model cannot replace and judgement.Fig2

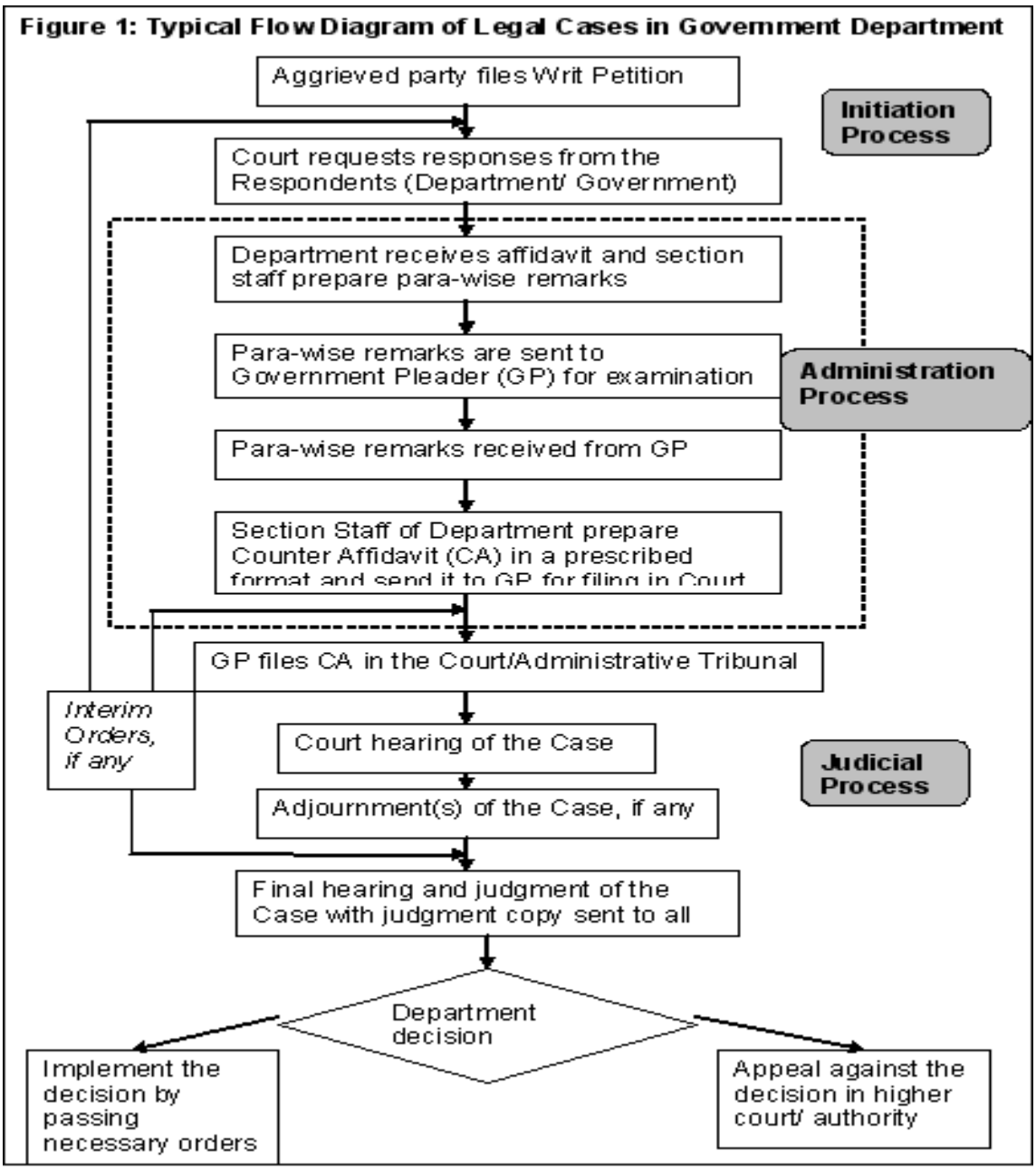

Fig 2 Organizational chart depicting proceedings within an agency.

Fig 2 - illustrates the legal case handling process in a government department. It begins with the Initiation Process, where an aggrieved party files a writ petition, prompting the court to request a response from the relevant department or government. This initiates the Administration Process, where the department receives the affidavit and its section staff prepare detailed para-wise remarks, which are then sent to the Government Pleader (GP) for examination. The GP reviews and returns the remarks, after which the department staff prepare a counter affidavit (CA) in a prescribed format and sends

it to the GP for filing in court. Once the CA is filed by the GP in court, the process moves to the Judicial Process. Here, court hearings are conducted, and there may be interim orders or adjournments. Eventually, the final hearing is held, and a judgment is issued. A copy of the judgment is sent to all concerned parties. Following the judgment, the department takes a decision. If the judgment is accepted, the department implements the necessary orders. Alternatively, if dissatisfied, the department can appeal the decision in a higher court or authority. This diagram captures the step-by-step legal procedures involved in managing legal cases in a government department.

4. Conclusion:

The proposed AI-based data extraction model is full of promise and serves as a proof of the possibility of making case scrutiny and removal of defects highly efficient through legal documents. Advanced NLP techniques combined with domain-specific training data ensure that relevant information, appropriate accuracy, precision, and recall are extracted by this model. The model's successful identification of petition formats, differentiation of legal provisions, and capability of capturing contextual features itself presents value to legal professionals. The advantages of the model are that the time is saved through the automated extraction of data, which saves the number of manual efforts and therefore there is saving of time, accuracy-the discrepancies and errors are reduced in the results of the extraction with the model, better decision-making data extracted helps in making the right decisions and strategic planning.

Reference:

1. Dr R. Naveen Kumar, Amit Kumar Bhore, Sourav Sadhukhan, Dr G. Manivasagam, Rubi Sarkar “Self-Monitoring System for Vision-Based Application Using Machine Learning Dr R Naveenkumar /Afr.J.Bio.Sc. 6(14) (2024) Page 11272 to 10 Algorithms” DOI 10.5281/zenodo.10547803, Vol 18 No 12 (2023), Page No 1958 – 1965, Published on 31-12-2023.
2. Dr. R .Naveenkumar “An Empirical Research Approach on Confusion Matrix Using Existing Musical Industry Dataset” International Journal of Scientific Research in Engineering and Management (IJSREM) Volume 08 Issue 04 | April - 2024 SJIF Rating 8.448 ISSN 2582-3930.
3. Amisha MP, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. J Fam Med Prim Car. 2019;8(5):2328–59.
4. González AM, Calatayud LR, Oliveira NG, Ustrell-Torrent JM. Artificial intelligence in orthodontics: Where are we now? A scoping review. Orthod Craniofac Res. 2021;24(2):6–15.
5. Abizadeh N, Moles DR, JO’Neill, Noar JH. Digital versus plaster study models: how accurate and reproducible are they? J Orthod. 2012;39(3):151–60.
6. Yılmaz H, Özlü FÇ, Karadeniz C, Karadeniz E. Time-Efficiency and Accuracy of Three-Dimensional Models Versus Dental Casts: A Clinical Study. Turk J Orthod. 2019;32(4):214–22.
7. Araújo TM, Caldas LD. Tooth extractions in Orthodontics: first or second premolars? Dental Press J Orthod. 2019;24(3):88–98.
8. Bogataj J, Gantar B. Pomenzobnihekstrakcij v celjustniortopediji. Role of tooth extractions in orthodontics. Zobo Drav Vestn. 1989;44(3):58–67.
9. Selvaraj M, Sennimalai K. Orthodontic model analysis in the permanent dentition: A review of past, and current methods. IP Ind J Orthod Dentofac Res. 2022;8(4):220–6.
10. Weintraub JA, Vig PS, Brown C, Kowalski CJ. The prevalence of orthodontic extractions. Am J Orthod Dentofac Orthop. 1989;96(6):462–6.
11. Rudge SJ, Jones PT, Hepenstal S, Bowden D, Cardiff C. The reliability of study model measurement in the evaluation of crowding. Euro J Ortho. 1983;5(1):225–31.
12. Battagel JM. The Assessment of Crowding Without the Need to Record Arch Perimeter. Part I: Arches With Acceptable Alignment. Brit J Orthod. 2016;23(2):137–44.
13. Somnath Mullick, R Naveenkumar, Sandip Bhattacharjee, Rahul Singha Applied Machine Learning for Predicting Crop Performance: A Supervised Learning Perspective 2024/8/14 Journal of Informatics Education and Research Volume 4 Issue 3 . DOI: https://doi.org/10.52783/jier.v4i3.1317