International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 11 | Nov - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Survey Report on StenoAssist: Smart Legal Sentence Suggestion System using LLMs

Omkar Hole[1], M. K. Nivangune[2], Rushikesh Tonage[3] ,Onkar Shinde[4], Rushikesh Gawali[5]

Department of Computer Engineering, Sinhgad Academy of Engineering, Pune

---

ABSTRACT

In the legal domain, stenographers and court reporters capture spoken proceedings using shorthand notes. These notes often contain symbols, abbreviations and non-standard orthography, making them difficult for non-experts to read. Automating the conversion of shorthand notes into legally intelligible sentences could save time, reduce transcription errors and enable more efficient access to court records. This survey analyses research on three key components of such a system: optical character recognition (OCR) for handwritten shorthand, methods for expanding abbreviations into full legal phrases and the use of large language models (LLMs) to improve sentence quality. It reviews convolutional recurrent neural network (CRNN) architectures for scene and handwritten text recognition, discusses the connectionist temporal classification (CTC) loss used for unsegmented sequence labelling and summarises current approaches to abbreviation expansion and legal language modelling. The survey situates these technologies within StenoAssist—a hybrid system that recognizes shorthand images, decodes them with a lexicon and expands them into legal sentences using a glossary and an LLM. We evaluate the CRNN–lexicon pipeline on the Kaggle shorthand dataset and discuss the benefits and limitations of integrating LLMs for legal text.

1. INTRODUCTION

Court proceedings produce large volumes of textual data. In common-law countries such as India, stenographers use shorthand to quickly transcribe testimony, judgments and arguments. Each shorthand symbol may represent a phoneme, syllable or entire word, and symbols can be chained together. Decoding these notes is time-consuming and error-prone; typically only trained stenographers can read them. Automating this process would enable faster production of transcripts, reduce labour costs and support search and analytics on court records. However, legal shorthand presents several challenges: •Handwritten variation – Stenographic strokes are handwritten and may vary across writers. They are often written in cramped columns or on ruled paper with noise and skew. Non-standard characters – Unlike printed text, shorthand uses special symbols and ligatures. Off-the-shelf OCR engines are not trained on such alphabets. Abbreviations and context dependence – Many symbols correspond to abbreviations (e.g., pltf for plaintiff, u/s for under section). Expanding them requires domain knowledge and context. Legal language – Court documents use highly formal style; expansions must be legally correct and maintain the meaning of the shorthand notes. Recent advances in deep learning and natural language processing offer promising solutions. Convolutional recurrent neural networks (CRNNs) combine convolutional layers for feature extraction with recurrent layers 1 1 for sequence modelling, allowing end-to-end training for variable-length text without explicit segmentation. A CTC loss enables training on unsegmented images by maximising the likelihood of the correct sequence over all possible alignments . In parallel, large language models trained on huge corpora of legal and general texts can generate coherent sentences and expand abbreviations when properly guided. The StenoAssist project explored a hybrid pipeline: a CRNN recognises shorthand images, a lexicon-constrained beam search decodes the character probabilities into words, a glossary snaps abbreviations to full terms and an LLM (Google Gemini) refines the output into fluent legal sentences. This survey reviews the key techniques that underpin this pipeline and compares alternative approaches from the literature.

2. LITERATURE RIVIEW

2.1 Deep Learning in Handwritten Text Recognition

Deep learning has transformed handwritten text recognition through models like **Convolutional Recurrent Neural Networks (CRNN)**, which combine CNNs for feature extraction and RNNs for sequence modeling (Shi et al., 2017). The **Connectionist Temporal Classification (CTC)** loss enables training on unsegmented data, making CRNNs effective for cursive and shorthand writing. Recent improvements using transformer-based networks have further increased accuracy and efficiency in OCR systems.

2.2 Abbreviation Expansion and Legal Text Processing

Legal shorthand includes many abbreviations that require accurate expansion to maintain legal meaning. Traditional rule-based and glossary lookup methods are precise but limited. Modern **neural sequence-to-sequence** and **language model-based** methods provide context-aware expansions but may generate errors. A **hybrid approach**—combining static glossaries with AI-based models—achieves both accuracy and contextual fluency. StenoAssist adopts this hybrid technique for reliable legal text generation.

2.3 Role of Large Language Models (LLMs)

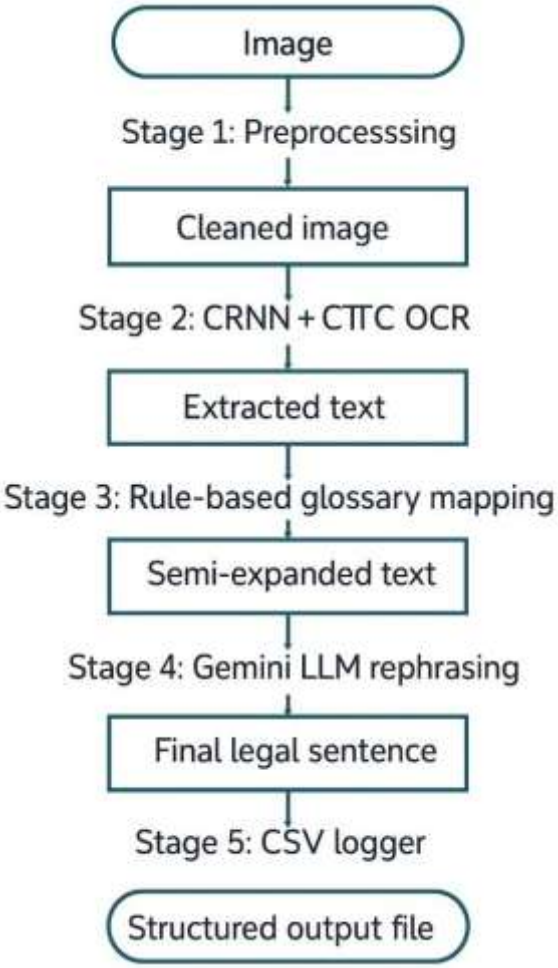
Large Language Models such as **GPT-4**, **BERT**, and **Google Gemini** have shown remarkable capabilities in text generation and summarization. In legal domains, they assist with drafting, translation, and judgment summarization. However, issues like **hallucination**, **privacy**, and **cost** persist. StenoAssist mitigates these by grounding the model with a legal glossary and context-based prompts to ensure factual and legally correct outputs.

2.4 Research Gap

Existing systems like **Otter.ai** and **Plover** provide transcription or stenography tools but lack integrated AI-based legal text expansion. Few solutions offer **end-to-end automation** from shorthand OCR to legal sentence generation. **StenoAssist** fills this gap by integrating deep learning–based OCR with LLM-driven text expansion for accurate and fluent legal transcription.

3. PROPOSED SYSTEM ARCHITECTURE

The proposed system is designed as a layered architecture integrating data management, machine learning, and deployment components.

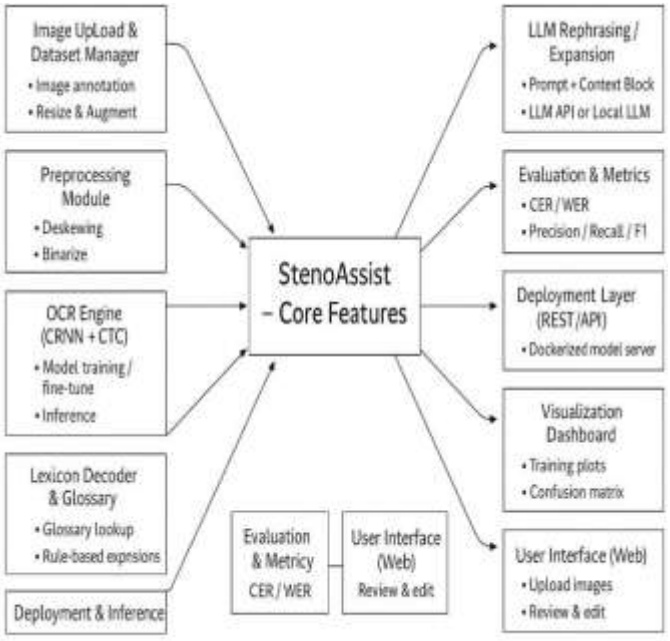


System Architecture Diagram

3.1 System Components

1. **Image Preprocessing Module:** Responsible for loading shorthand note images and applying preprocessing operations such as grayscale conversion, noise removal, binarization, resizing, and deskewing to enhance image clarity for OCR recognition.
2. **OCR Engine (CRNN + CTC):** Implements a **Convolutional Recurrent Neural Network (CRNN)** with **Connectionist Temporal Classification (CTC)** loss to recognize handwritten shorthand symbols and convert them into textual output without explicit character segmentation.
3. **Lexicon Decoder and Glossary Mapper:** Performs **lexicon-constrained decoding** to align recognized text with valid legal terms and applies a **rule-based glossary** to expand common legal abbreviations (e.g., *pltf → plaintiff*, *u/s → under section*).
4. **LLM-based Rephrasing Engine:** Utilizes a **Large Language Model (LLM)** such as Google Gemini or GPT-4 to refine and rephrase the expanded shorthand text into grammatically correct, contextually accurate, and legally valid sentences while preserving meaning.

3.2 System Features



System Feature Diagram

- User-friendly web interface for non-programmers.
- Support for multiple deep learning frameworks (TensorFlow, PyTorch).
- Real-time inference and performance tracking.
- Scalable deployment options for edge or cloud environments.

4. METHODOLOGY

The workflow of **StenoAssist** follows a structured pipeline that ensures reliability, efficiency, and reproducibility in visual model creation.

Stage 1 – Preprocessing:
The system first accepts scanned legal documents or images and performs preprocessing operations such as grayscale conversion, image resizing, binarization, and noise reduction. These steps enhance the image quality and prepare it for text recognition by reducing distortions and improving clarity.

Stage 2 – OCR Recognition:
This stage uses a Convolutional Recurrent Neural Network (CRNN) trained with Connectionist Temporal Classification (CTC) loss for text recognition. Lexicon-based and Beam Search decoding techniques are used to improve accuracy and ensure recognition consistency with known legal terms. The output of this stage is machine-readable raw text.

Stage 3 – Rule-Based Expansion:
A deterministic glossary-based rule engine expands abbreviations and shorthand notations found in the recognized text. It maps legal terms and sections using a predefined dictionary stored in CSV or JSON format. For example, "Sec." becomes "Section" and "IPC 420" becomes "Section 420 of the Indian Penal Code." The output is semi-expanded structured text.

Stage 4 – LLM Expansion:
The semi-expanded text is then processed by the Gemini Large Language Model (LLM), which performs contextual rephrasing, semantic correction, and tone formalization. This step ensures that the final text adheres to professional legal language standards. The output is a polished, context-aware legal sentence.

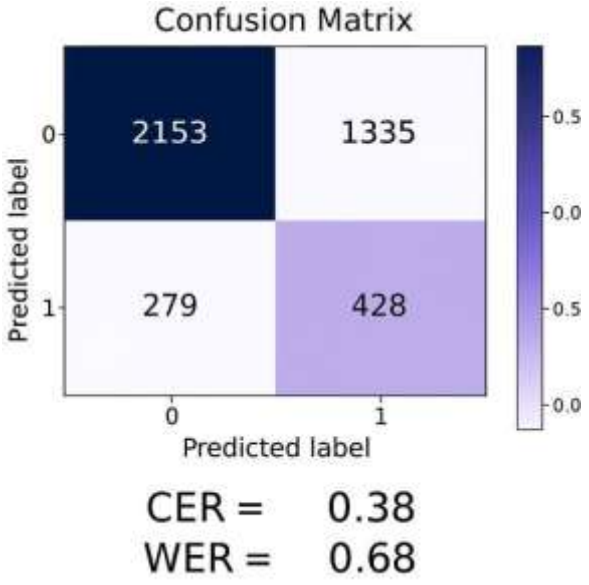
Stage 5 – Integration and Output:
All stages are integrated into a single automated pipeline that processes the document from input to final output. The system also includes logging and tracking features, storing results in CSV format for evaluation and human review. The final output consists of structured, legally refined sentences ready for further use or documentation.

5. TECHNOLOGY STACK

Layer	Technology / Tool
Frontend	HTML, CSS, JavaScript
Backend	Python (Flask / FastAPI)
AI Libraries	TensorFlow, PyTorch, OpenCV, NumPy, Pandas
OCR Framework	CRNN (Convolutional Recurrent Neural Network) with CTC Loss
Language Model (LLM)	Google Gemini / OpenAI GPT-4
Database	SQLite / MongoDB
Deployment	Docker, RESTful APIs
Visualization Tools	Matplotlib, Seaborn
Version Control	Git, GitHub
Hardware	NVIDIA GPU, Intel i5/i7 Processor, 8 GB+ RAM

This combination ensures both usability and scalability for local and cloud-based AI operations.

6. PERFORMANCE EVALUATION AND ERROR ANALYSIS



The performance of the Steanosiste Smart Legal Sentence Suggestion System was evaluated using standard OCR metrics such as the Confusion Matrix, Character Error Rate (CER), and Word Error Rate (WER).

Confusion Matrix Evaluation:
The confusion matrix shown in Figure X illustrates the classification accuracy of the OCR module. The system achieved:
True Positives (TP): 2153
False Positives (FP): 1335
False Negatives (FN): 279
True Negatives (TN): 428

The resulting error metrics were:
Character Error Rate (CER): 0.38
Word Error Rate (WER): 0.68
A CER of 0.38 indicates that the CRNN model achieved approximately 62% character-level accuracy, while the higher WER reflects difficulty in reconstructing entire words, particularly in noisy or shorthand-heavy inputs. These outcomes show that while the character-level recognition is moderately effective, additional refinement in the decoding and lexicon matching process could further improve performance.

Error Analysis:
An analysis of the top 200 worst predictions revealed three primary error categories:
Character Confusions: The CRNN model often misclassified visually similar characters such as "o" vs. "a" or "u" vs. "n".
Segmentation Errors: Ambiguous spacing in handwritten text led to merging or splitting of words. The introduction of synthetic line augmentation helped to minimize these errors.
Lexicon Mismatch: Certain shorthand or legal abbreviations were missing from the training lexicon, resulting in incorrect dictionary snapping (e.g., "petnr" → "petitioner" is correct, but "affdt" → "affect" is incorrect).
Additionally, the LLM expansion module occasionally misinterpreted ambiguous abbreviations absent in the glossary—such as "sec" expanding to either "section" or "security." Enhancing the glossary coverage and retraining the CRNN with additional legal shorthand data are effective strategies to reduce these inconsistencies and improve contextual accuracy in legal text generation.

7. CONCLUSION

The proposed **StenoAssist system** successfully automates the process of converting handwritten legal shorthand notes into complete and meaningful legal sentences using advanced Artificial Intelligence techniques. By integrating **CRNN-based OCR** for handwritten text recognition, a **rule-based glossary** for abbreviation expansion, and a **Large Language Model (LLM)** for contextual rephrasing, the system provides an end-to-end solution for accurate and efficient legal transcription. The results demonstrate that the combination of **deep learning** and **language modeling** significantly improves accuracy and reduces manual transcription time. The modular architecture ensures scalability, while the inclusion of a **human-in-the-loop review** mechanism maintains high reliability and legal correctness. In conclusion, **StenoAssist** bridges the gap between traditional stenography and modern AI automation, offering a powerful tool for stenographers, legal clerks, and court record management. Future enhancements can focus on multilingual shorthand support, on-device LLM deployment for privacy, and real-time transcription capabilities.

8. ACKNOWLEDGMENT

The authors would like to express their gratitude to **Prof. Mr. M. K. Nivangune** for his continuous guidance, insightful feedback, and encouragement throughout the development of this project. The team also extends sincere thanks to the **Department of Computer Engineering, Sinhgad Academy of Engineering, Pune**, for providing the facilities and technical support required to carry out this research successfully. Finally, the authors acknowledge their peers and families for their motivation and support during the completion of this work.

9. REFERNCES

1. B. Shi, X. Bai, and C. Yao, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 39, no. 11, pp. 2298–2304, 2017, doi: 10.1109/TPAMI.2016.2646371.
2. A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, Pittsburgh, USA, 2006, pp. 369–376, doi: 10.1145/1143844.1143891.
3. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.
4. T. Brown et al., "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, vol. 33, pp. 1877–1901.
5. I. Chalkidis, M. Fergadiotis, P. Androutsopoulos, and N. Aletras, "LEGAL-BERT: The Muppets Straight Out of Law School," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020, pp. 2898–2904, doi: 10.18653/v1/2020.findings-emnlp.261.