

IntelliVision: An AI-Based FIR Draft Generation System

Krupali Dhawale, Payal Chadhokar, Prarthana Shukla, Sakshi Baghel, Vishakha Varani

Department of Artificial Intelligence, G. H. Raisoni College of Engineering, Nagpur, India

Published October 2026

Abstract

IntelliVision is an AI-based FIR draft generation system that processes text, speech, and image-based complaints using OCR, speech recognition, and NLP. It extracts key information, classifies crime types, retrieves similar cases, and generates an AI-assisted FIR draft for officer review. The crime classification module uses TF-IDF and Logistic Regression and achieved 100% test accuracy on a synthetic dataset of 515 complaints. The system aims to reduce documentation effort and improve the consistency of FIR preparation.

Keywords: artificial intelligence; FIR draft generation; legal NLP; named entity recognition; OCR; speech recognition; information extraction; large language models; human-in-the-loop

1 Introduction

First Information Reports (FIRs) are an important part of the criminal justice process and require accurate recording of information provided by complainants. Preparing an FIR draft manually can be time-consuming, particularly when complaints contain unstructured text, spoken statements, or image-based documents. Important details may also be missed or inconsistently organized during manual documentation.

IntelliVision is an AI-based FIR draft generation system designed to assist officers in converting multimodal complaints into structured information. The system uses Optical Character Recognition (OCR), speech recognition, Natural Language Processing (NLP), and Named Entity Recognition (NER) to extract relevant details from complaints. It also performs crime-type classification and similar-case retrieval before generating an AI-assisted FIR draft.

The system uses TF-IDF and Logistic Regression for crime classification and incorporates an agentic workflow to coordinate different processing modules. The generated FIR is intended as a draft that can be reviewed and modified by an authorized officer. Thus, IntelliVision focuses on reducing documentation effort while maintaining human oversight and supporting consistent FIR preparation.

2 Objectives and problem statement

The objectives are to accept text, voice, image, and scanned complaints; normalize them into a common representation; extract complaint-specific entities; provide an advisory crime category; retrieve related legal or case information; identify missing and conflicting details; coordinate the stages through an agentic workflow; generate a structured FIR draft; and require authorized human inspection before final use.

The central problem is the conversion of heterogeneous complaint information into a structured FIR draft without introducing unsupported facts. Manual preparation requires repeated reading, transcription, entity identification, classification, information retrieval, and formatting. Direct LLM generation can conceal missing information or produce plausible but unsupported details. IntelliVision addresses this problem through validation before generation and by instructing the drafting model to leave unavailable fields unspecified rather than guessing.

The system also treats each component as fallible. Speech recognition, OCR, NER, classification, retrieval, and generation have different failure modes. Conditional routing and stored intermediate outputs allow the investigation agent to request additional processing or human intervention when required.

3 Related work

Legal information extraction has developed around named entity recognition, relation extraction, event detection, and document classification. Premasiri et al. surveyed legal information extraction and identified challenges involving datasets, domain variation, multilingual text, and generalization [1]. Police-report element extraction using LLMs has also been investigated [2], while fine-grained legal NER demonstrates the importance of domain-specific entity categories [3]. Structured attribute extraction from criminal-case documents provides another direction for legal LLM applications [4].

Transformer language models provide strong contextual representations for NLP [10], and legal-domain pretraining has been investigated for Indian law [11]. LegalBench provides a benchmark for evaluating different forms of legal reasoning in LLMs [12]. These studies motivate domain-aware extraction and evaluation rather than assuming that general-purpose models are sufficient.

For knowledge grounding, Lewis et al. introduced retrieval-augmented generation, which combines parametric generation with retrieved external information [6]. ReAct further demonstrated a reasoning-and-action paradigm in which a language model interacts with external tools [13]. IntelliVision adopts the underlying idea of controlled tool use but constrains retrieved legal information to supporting context that remains distinguishable from the original complaint.

Speech recognition is relevant to voice complaints. Radford et al. reported robust multilingual speech recognition from large-scale weak supervision [7], while wav2vec 2.0 showed that self-supervised speech representations can reduce dependence on labeled audio [8]. For document images, transformer-based OCR such as TrOCR provides an alternative to conventional OCR pipelines [9]. These developments support the multimodal design of IntelliVision.

Additional foundations relevant to the implementation include expert-annotated legal datasets, legal-language benchmarks, Legal-BERT, CaseHOLD, document-layout modeling, contextual representations, neural NLP foundations, word representations, information retrieval, classical text classification, scikit-learn, spaCy, NLTK, Transformers, retrieval and long-context generation methods, and Llama-family instruction-tuned models.

Table 1. Comparison with representative related work.

Study	Task	Technique	Human Review
Premasiri et al. [1]	Legal IE	Survey	Not applicable
Xing & Chen [2]	Police reports	LLM extraction	Not specified

Study	Task	Technique	Human Review
Leitner et al. [3]	Legal NER	Fine-grained NER	No
Paul et al. [11]	Indian legal NLP	Legal PLMs	No
IntelliVision	Multimodal FIR draft	Agentic pipeline	Mandatory

The literature also indicates that legal NLP differs from general-domain NLP because terminology, document structure, citation patterns, and jurisdiction-specific concepts affect model behavior. LegalBERT and related domain-adapted models show that continued pretraining on legal corpora can improve downstream performance, while Indian-law studies indicate that local legal language and terminology deserve specific treatment [11]. For IntelliVision, a future production version should therefore combine general language capability with controlled legal resources, domain-specific extraction models, and explicit provenance for retrieved information. LegalBench demonstrates that legal reasoning is multidimensional rather than a single classification task [12]. Document-understanding research likewise shows that recognition quality can depend strongly on layout, scan quality, and document type [9]. Consequently, the proposed system evaluates individual modules separately before considering end-to-end workflow quality.

4 Proposed architecture

IntelliVision follows a staged architecture consisting of multimodal input processing, normalization, NLP extraction, validation, classification, retrieval, agent-state management, FIR generation, and human review. Text complaints enter preprocessing directly; speech is converted to text; and photographs or scans are processed through OCR. The normalized representation is then passed to common downstream components.

The validation layer checks required fields and possible contradictions. Crime classification provides an advisory category, while similar-case retrieval searches an available repository. The generation stage receives structured complaint fields, validation results, classification output, and clearly separated retrieval context. The LLM is instructed to generate only from validated information.

The architecture is designed for traceability. Rule-based checks and model-based language generation remain separate, and each module records its output and status. This separation supports diagnosis when an error is found and reduces the chance that an unsupported generated statement is silently accepted.

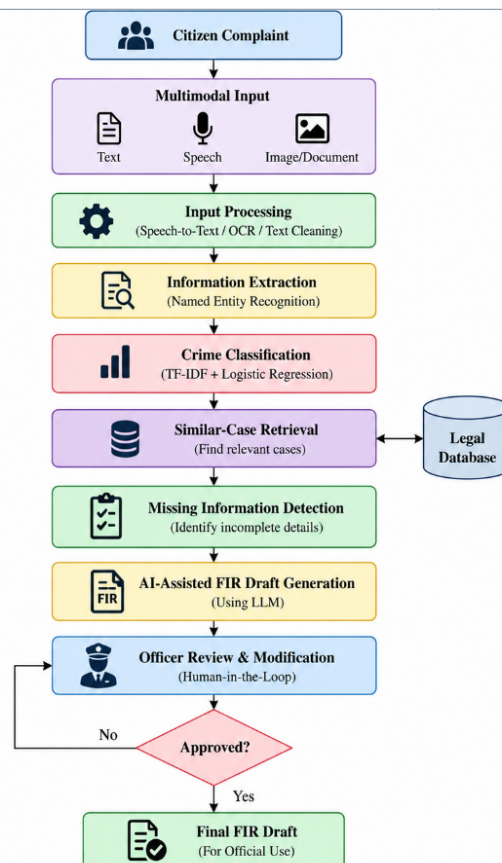


Fig. 1. Workflow of the Proposed IntelliVision System.

Figure 1.

The architecture can be viewed as four logical layers. The input layer handles modality-specific acquisition and normalization. The intelligence layer performs preprocessing, NER, classification, retrieval, and validation. The orchestration layer maintains workflow state and decides which operation should be executed next. The presentation layer provides the draft, warnings, source information, and approval controls to the authorized reviewer. This separation reduces coupling between components and allows individual modules to be replaced without redesigning the complete system. A key design principle is provenance preservation: each extracted field should retain a source reference indicating whether it originated from typed text, a speech transcript, OCR output, or a structured database. Reviewer modifications should also be recorded. The architecture further distinguishes deterministic checks from generative operations. Required-field checks, data-format checks, and consistency rules can be performed deterministically, while summarization and natural-language drafting are handled by the LLM.

5 Multimodal input and NLP information extraction

Every complaint is converted to a common textual representation while preserving the original modality. Typed complaints are normalized and segmented into sentences. Voice complaints pass through speech recognition, while images and scanned documents pass through OCR. The original audio or image remains associated with the extracted text for later verification.

Speech recognition is treated as an interpretation rather than absolute ground truth because accents, noise, multiple speakers, and domain-specific terms can affect transcription. OCR can be affected by

blur, skew, low resolution, handwriting, and complex layouts. The prototype uses EasyOCR for image processing; transformer-based OCR such as TrOCR provides a potential future alternative [9].

The NLP stage uses preprocessing and NER to identify entities such as complainants, accused or involved persons, locations, dates, times, objects, vehicles, monetary amounts, witnesses, and evidence references. The structured record may contain Incident_Date, Incident_Time, Location, Persons, Event_Description, Objects, Witnesses, and Supporting_Evidence. Information that is not explicitly present is marked missing rather than estimated.

This distinction between extraction and inference is important for legal applications. General language models can produce fluent completions, but legal-domain models and benchmarks show that domain-specific evaluation remains necessary [11], [12].

After modality-specific conversion, the normalized complaint is represented as a structured intermediate record containing the original text, source modality, extracted entities, confidence values where available, missing fields, contradiction flags, and source references. Keeping this intermediate representation prevents later stages from repeatedly interpreting raw input and provides a stable interface between modules. NER output must also be interpreted carefully. An extracted person name is not automatically an accused person, and a location mentioned in a narrative is not necessarily the incident location. Future versions can combine NER with relation and event extraction so that entities are linked to their roles and events, consistent with legal information extraction research [1], [3]. Missing-information detection remains a validation task rather than a generation task. If a required field is absent, the system records a missing-field flag and can request clarification instead of allowing the LLM to infer a value.

6 Crime classification and similar-case retrieval

Crime classification is implemented as an assistive supervised learning task using TF-IDF features and Logistic Regression. TF-IDF represents the importance of a term in a complaint relative to the corpus. For a term t in document d , the feature can be expressed as

$$\text{TFIDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

$$\text{IDF}(t) = \log \left(\frac{N}{df(t)} \right)$$

The classifier predicts an advisory crime category and can provide a confidence value. The prediction is not treated as a legal determination. Low-confidence or ambiguous cases can be routed for clarification or human review.

Similar-case retrieval searches an available repository for records sharing relevant characteristics with the current complaint. Retrieved information is treated as reviewer context rather than evidence. RAG provides a general architecture for combining retrieval and generation [6], while legal information retrieval research highlights the importance of domain-specific indexing and evaluation. Precision@K and Recall@K are suitable retrieval measures.

For classification, the principal evaluation measures are:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

These formulas make the reported results reproducible and distinguish overall correctness from class-specific performance.

The TF-IDF representation provides a transparent baseline for complaint classification because important words and phrases contribute directly to the feature vector. Logistic Regression then learns class-specific weights from the training data. This approach is computationally lightweight and suitable for a prototype, although a larger deployment may require domain-adapted transformer models [10], [11]. The present experiment deliberately uses a conventional classifier so that the relationship between input text, features, and predicted category can be examined clearly. A confusion matrix is useful when several crime categories have similar language. Future experiments should report class distributions, macro-averaged metrics, and confidence calibration. Retrieval should similarly be evaluated independently from generation. A relevant case appearing in the top five results is a retrieval success even if the LLM later summarizes it incorrectly; fluent generation cannot compensate for irrelevant retrieval.

7 Agentic workflow and FIR draft generation

The investigation agent coordinates the stages through conditional routing. The workflow begins with modality detection: text proceeds to preprocessing, speech proceeds through speech recognition, and image or scan input proceeds through OCR. All routes then converge on preprocessing, NER, structured extraction, and validation.

The agent maintains separate state for extracted facts, warnings, classification results, and retrieval context. This design is related to tool-using LLM paradigms such as ReAct [13], but IntelliVision constrains the workflow around factual traceability and mandatory human review. If required information is missing or conflicting, the workflow can flag the issue before drafting.

FIR generation begins only after validation. The LLM receives structured fields and controlled formatting instructions. It is explicitly instructed not to invent facts and to leave unavailable information unspecified. Retrieved legal or case information is not merged with complaint facts. After generation, an authorized officer reviews the draft, corrects it where necessary, and approves or rejects it. ReportLab can then be used to produce a PDF representation of the approved draft.

The agentic workflow is designed around controlled state transitions rather than unrestricted autonomous behavior. A workflow state can contain the input modality, normalized text, extracted entities, validation flags, predicted category, retrieved context, and draft status. Each node reads the state, performs one controlled operation, and writes its result back. Conditional edges can route incomplete or low-confidence cases to clarification or review. This improves observability: if a generated field is incorrect, the reviewer can trace it backward to the structured record and then to the source modality. If the source text was misrecognized, the problem can be routed to OCR or speech recognition. The generation prompt should explicitly define the system role, permitted information, output structure, and prohibited behavior. The model should not create dates, names, legal sections, evidence, witness statements, or event details that are absent from the validated state.

8 Implementation, dataset and evaluation

The prototype is implemented using Python and Flask with an HTML, CSS, Bootstrap, and JavaScript interface. MySQL stores structured records. spaCy and NLTK support preprocessing and NER, scikit-learn provides classification, EasyOCR supports image processing, a speech-recognition component

supports voice input, Llama 3 is used for controlled text generation, LangGraph manages workflow state, ReportLab supports PDF generation, and GitHub is used for version control.

The crime-classification experiment uses a synthetic dataset of 515 complaint records. An 80:20 split gives 412 training records and 103 test records. Complaint text is transformed using TF-IDF and classified using Logistic Regression. Because the dataset is synthetic, the experiment measures prototype performance rather than real-world police deployment performance.

Speech recognition can be evaluated with Word Error Rate (WER) and Character Error Rate (CER):

$$CER = \frac{S_c + D_c + I_c}{N_c}$$

where S, D, and I denote substitutions, deletions, and insertions, and N is the number of reference words. Retrieval can be evaluated using:

$$\text{Precision @ } K = \frac{\text{Relevant results in top } K}{K}$$

$$\text{Recall @ } K = \frac{\text{Relevant results in top } K}{\text{Total relevant results}}$$

The OCR and speech modules were functionally tested, but quantitative WER, CER, and field-level accuracy were not calculated because annotated ground-truth datasets were unavailable.

The evaluation design separates component-level performance from end-to-end functionality. Classification is measured quantitatively because a labeled synthetic dataset is available, whereas OCR and speech recognition are reported as functional tests because no ground-truth benchmark was prepared. The 80:20 split provides 412 training samples and 103 test samples. Although the resulting test performance is 100%, the dataset size and synthetic construction mean that the estimate may be optimistic. A future evaluation should use an independently collected test set, stratified class distributions, repeated splits or cross-validation, and preferably an external dataset. For end-to-end evaluation, reviewers can score whether each required field is preserved from input to final draft, whether missing information remains marked as missing, whether retrieved context is correctly separated, and whether human corrections are reflected in the final output. A field-level measure can be defined as:

$$\text{Field Preservation} = \frac{\text{Correctly preserved required fields}}{\text{Total required fields}}$$

9 Results, security, limitations and future scope

The crime-classification model achieved 100% accuracy, precision, recall, and F1-score on the 103-sample held-out test set. Manual testing on 10 complaint examples produced 9 correct classifications, corresponding to 90% accuracy. One online-harassment example was classified as extortion, illustrating ambiguity between semantically related complaint categories. The results should therefore be interpreted only in the context of the synthetic dataset.

The complete prototype workflow was functionally tested from complaint submission through text processing, classification, legal information retrieval, FIR generation, and officer review. OCR and speech recognition successfully produced text for subsequent processing, but standardized quantitative benchmarking remains future work.

Because an FIR-support system may process names, contact information, locations, and descriptions of alleged offences, security controls are required. Authentication, role-based access, protected storage, controlled media access, audit logging, and secure retention and deletion should be used. The interface should distinguish extracted facts, retrieved information, generated text, and officer modifications.

The main limitations are the synthetic dataset, limited evaluation size, modality-specific errors, dependence on repository quality, possible LLM hallucination, and the absence of quantitative OCR and speech benchmarks. Future work should add multilingual support for Indian languages, larger authorized datasets, domain-adapted legal models, improved OCR, confidence-aware validation, guided clarification, retrieval from approved legal sources, robustness testing, and controlled integration with authorized systems. Indian legal-domain pretraining has already shown the value of adapting language models to local legal text [11], providing a useful direction for future development.

Table 2. Functional modules.

Module	Function	Output
Input Processing	Accept text, voice, image, or scan	Normalized input
OCR / Speech	Convert image/audio to text	Extracted text
NLP / NER	Extract complaint entities	Structured record
Validation	Detect missing/conflicting fields	Warnings
Classification	Suggest crime category	Predicted class
Retrieval	Find related information	Case/legal context
FIR Generation	Draft from validated data	Reviewable FIR

The observed classification result should not be interpreted as evidence that IntelliVision can determine the legal section of an offence automatically. Crime labels in the synthetic dataset represent an experimental prediction target; actual legal characterization may require facts, evidence, statutory interpretation, jurisdictional context, and authorized professional judgment. Security should be considered at both application and model levels. Uploaded images and audio can contain sensitive information, and prompt or retrieval pipelines can expose information if access controls are weak. Role-based access can restrict who can view complaints, edit records, approve drafts, or export documents. Audit logs should record these actions without unnecessarily duplicating sensitive content. Prompt injection and untrusted retrieved content are additional concerns for future versions. Retrieved documents should be treated as data rather than instructions to the LLM, and trusted workflow instructions should remain separate from external text. Future validation should include adversarial and incomplete complaints, multilingual inputs, noisy speech, poor-quality scans, contradictory statements, and ambiguous crime descriptions.

10 Conclusion

IntelliVision presents an agentic AI framework for assisting FIR draft preparation from multimodal complaints. The proposed workflow combines speech recognition, OCR, NLP, NER, validation, crime classification, retrieval, LLM-based drafting, and mandatory human review within a traceable pipeline.

The classification component achieved 100% performance on the held-out synthetic test set and 90% accuracy during manual testing. These results demonstrate prototype feasibility but do not establish real-world police performance. The absence of annotated datasets also prevents quantitative benchmarking of the OCR and speech modules.

The central contribution is the integration of multimodal processing, structured extraction, validation, retrieval, and controlled generation while preserving human responsibility for the final FIR. Future evaluation on larger, diverse, and appropriately authorized datasets is necessary before deployment in operational environments.

The proposed framework is intended to reduce repetitive documentation work while preserving factual and procedural control. Its modular design allows improvements to OCR, speech recognition, legal NLP, retrieval, classification, or generation without requiring the entire system to be replaced. The most important deployment requirement is that automation remains assistive: the system can organize information and prepare a draft, but an authorized officer remains responsible for verification and approval.

Acknowledgments

The authors thank the Department of Artificial Intelligence, G. H. Raisoni College of Engineering, Nagpur, for its guidance and support throughout this project.

References

- [1] Premasiri D, Ranasinghe T, Mitkov R, El-Haj M, Frommholz I. Survey on legal information extraction: Current status and open challenges. *Knowl Inf Syst.* 2025;67(12):11287–11358. <https://doi.org/10.1007/s10115-025-02600-5>
- [2] Xing X, Chen P. Entity extraction of key elements in 110 police reports based on large language models. *Appl Sci.* 2024;14(17):7819. <https://doi.org/10.3390/app14177819>
- [3] Leitner E, Rehm G, Moreno-Schneider J. Fine-grained named entity recognition in legal documents. In: *Semantic Systems: The Power of AI and Knowledge Graphs*. Cham: Springer; 2019. p. 272–287. (Lecture Notes in Computer Science; vol. 11702). https://doi.org/10.1007/978-3-030-33220-4_20
- [4] Adhikary S, Sen P, Roy D, Ghosh K. A case study for automated attribute extraction from legal documents using large language models. *Artif Intell Law.* 2026;34:245–266. <https://doi.org/10.1007/s10506-024-09425-7>
- [5] Brown TB, et al. Language models are few-shot learners. In: *Proc. Adv. Neural Inf. Process. Syst.* 2020. p. 1877–1901.
- [6] Lewis P, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Proc. Adv. Neural Inf. Process. Syst.* 2020. p. 9459–9474.
- [7] Radford A, et al. Robust speech recognition via large-scale weak supervision. In: *Proc. 40th Int. Conf. Mach. Learn.*, PMLR. 2023. p. 28492–28518.
- [8] Baeovski A, Zhou H, Mohamed A, Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In: *Proc. Adv. Neural Inf. Process. Syst.* 2020. p. 12449–12460.
- [9] Li M, et al. TrOCR: Transformer-based optical character recognition with pre-trained models. In: *Proc. AAAI Conf. Artif. Intell.* 2023. p. 13094–13102. <https://doi.org/10.1609/aaai.v37i11.26538>
- [10] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proc. NAACL-HLT.* 2019. p. 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- [11] Paul S, Mandal A, Goyal P, Ghosh S. Pre-trained language models for the legal domain: A case study on Indian law. In: *Proc. 19th Int. Conf. Artif. Intell. Law.* 2023. p. 187–196. <https://doi.org/10.1145/3594536.3595165>
- [12] Guha N, et al. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In: *Proc. Adv. Neural Inf. Process. Syst.* 2023.

-
- [13] Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, et al. ReAct: Synergizing reasoning and acting in language models. In: *Proc. Int. Conf. Learn. Representations*. 2023.