Multimodal AI-Powered Incident Reporting: Leveraging LLMs for Structured, Accessible, and Legally-Usable Reports

Nisarga Dey
Department of Computer Engineering
Marwadi University, Rajkot, Gujarat, India

Om Prakash Suthar
Department of Computer Engineering
Marwadi University, Rajkot, Gujarat, India

Suraj Yaligar
Department of Computer Engineering
Marwadi University, Rajkot, Gujarat, India

Pratikkumar Dungano
Department of Computer Engineering
Marwadi University, Rajkot, Gujarat, India

Abstract- Underreporting and low-quality evidence are the results of traditional crime reporting systems, which are frequently slow, manual, and inaccessible. This paper describes a multimodal AI-powered incident reporting system that uses Large Language Models (LLMs) to transform unstructured data submitted by citizens into reports that are legally usable. With support for anonymous submissions to lessen user reluctance, the suggested framework allows real-time reporting via a mobile application utilizing text, images, videos, audio, and GPS metadata. To automatically extract, validate, and synthesize evidence, an integrated AI pipeline integrates speech-to-text, vision-language analysis, and cross-modal reasoning. The system facilitates offline reporting via peer-to-peer mesh networking to guarantee dependability in the event of network failures or disasters. Heatmaps, case management, and real-time crime visualization are all provided by a police-facing analytics dashboard. When compared to conventional reporting methods, the results show increased operational efficiency, accessibility, and evidence quality.

Keywords – Large Language Model, Artificial Intelligence, Legal Aid, Natural Language Processing, Retrieval Augmented Generation, Sustainable Development Growth.

I. INTRODUCTION

Public safety depends on crime reporting, but conventional reporting methods are still cumbersome, slow, and primarily manual. Fear of identification, a lack of confidence in confidentiality, and the inconvenience of physical or form-based reporting procedures are the main reasons why many crimes go unreported rather than because they go unnoticed. Due to these restrictions, law enforcement operations are less effective, responses are delayed, and evidence is of low quality. Despite the fact that contemporary smartphones have cameras, microphones, and GPS sensors that can record rich, real-time evidence, the majority of current crime reporting systems are unable to make effective use of this multimodal data. Because reports are usually unstructured and text-based, authorities must manually interpret them, which frequently results in errors and the loss of important contextual information. This gap creates the need to have intelligent systems that can automatically process and validate different forms of evidence. The current paper presents a Multimodal AI-Powered Incident Reporting solution that will provide mobile application to facilitate real-time, anonymous, and offline reporting [1]. The system enhances accessibility, quality of evidence, and operational effectiveness through converting the unstructured multimodal inputs into structured reports that can be used in the court of law and providing a real-time analytics and case management dashboard to law enforcement.

1.1 Background and Motivation

Digital governance has not replaced manual, text-based processes of crime reporting systems which are not encouraging citizen involvement and sluggish in law enforcement response. The evidence that is readily available through the modern smartphones in a multimodal format is not sufficiently used by the current platforms, which results in incomplete and legally questionable reports. Accessibility is also limited because of the fear of recognition and a dependence on total internet access [2]. It is based on these challenges that this piece of work examines the role of large language models and multimodal AI to automate evidence interpretation, enable anonymous and offline reporting, and offer a reliable, real-time reporting system to bridging the communication divide between citizens and authorities [3].

1.2 Problem Statement

Most of the existing crime reporting systems are text based and manual in nature and require 24-hour internet connectivity to automatically generate low quality evidence, slow response time and underreport. Fear of being identified, mistrust of confidentiality, and procedural inconvenience are common reasons why citizens are reluctant to report incidents. Rich multimodal evidence, including images, videos, audio, and location data, can be captured by smartphones; however, current platforms are unable to process and validate this data in a way that is both structured and legally usable. As a result, law enforcement organizations must interpret unstructured reports, which results in errors and inefficiencies. A secure, resilient reporting system that can convert multimodal inputs into standardized, actionable incident reports in real time is obviously needed [4][5].

1.3 Research Objectives

Through the following goals, this study seeks to address the aforementioned challenges:

- To design a multimodal incident reporting framework that enables citizens to submit real-time reports using text, images, videos, audio, and GPS metadata through a unified mobile platform.
- To leverage Large Language Models and multimodal AI techniques for automatically extracting, validating, and synthesizing unstructured evidence into structured, legally-usable incident reports.
- To ensure accessibility and resilience in reporting by supporting anonymous submissions and offline communication through peer-to-peer mesh networking during network outages or disaster scenarios.
- Most of the existing crime reporting systems are text based and manual in nature and require 24-hour internet connectivity to automatically generate low quality evidence, slow response time and underreport.

1.4 Contributions & Novelty

The study introduces the innovative end-to-end multimodal artificial intelligence-driven incident reporting system to balance citizen-generated information with law enforcement requirements. The primary contribution consists of automatic multimodal data to structured and legally useful reports, which is achieved through the joint operation of large language models and audio and visual processing [6]. The proposed structure will combine real-time reporting, cross-modal validation of evidences, anonymity, and offline communication resilient to disasters into a single platform, unlike the existing systems, which focus exclusively on reporting or visualization. In addition, situational awareness and data-driven policing are possible due to the use of the real-time analytics dashboard of the system, which makes it different compared to the old ways criminal reporting systems in terms of technology and operation.

1.5 Overview of Upcoming Sections

This is the way the remaining part of the paper is arranged. The state of AI-based methods and crime coverage systems are discussed in the Related Work section and reveal the disadvantages of these tools along with the gap in the research that the given study closes. The Methodology section entails an elaborate discussion of the proposed system architecture, multimodal data processing pipeline and AI-based report generation workflow. The subsequent one is System Implementation and Outcomes which is about the development of the prototype, the main features and the results seen. In the Conclusion section, the contributions are summed up, the effectiveness of the recommended approach is assessed, and recommendations about the future developments are provided.

II. RELATED WORK

The essentials of the present crime reporting systems are GIS- friendly maps, mobile/internet form entries, and straightforward information sharing of the citizens and the police. Even though they are more accessible, they do not offer automated interpretation of the evidence and are largely based on a manual and text-driven reporting. Most of the solutions lack an end-to-end multimodal processing, legal framework, anonymity, offline resilience in a single

reporting structure, and thus leave huge gaps in their solutions, which this study tries to address. The recent research explores the multimodal data fusion and AI-assisted intelligence report [7].

2.1 Literature Review

The reviewed literature shows that there are significant attempts, which are being made towards GIS-enabled visualization, digital and mobile-based crime reporting systems, and information sharing between citizens and law enforcement. Some of them focus on AI and multimodal data integration in intelligence and incident detection, but others focus on improving the access, anonymity and speed of reporting via mobile apps. The research gap that the work will fill can be explained by the fact that the existing solutions mainly predicate on manual or semi-automated procedures, are not based on solid multimodal reasoning and do not generate structured and legally-usable reports, as summarized in the comparative tables provided.

Table -1 Comparison of the Studies

Ref No.	Tech Stack Used	Key Findings	Merits	Demerits
[1]	Android, Java, PHP, MySQL, Maps API, ML (SVM, K-means)	SVM is 89.5% accurate for classification; K-means/DBScan find hotspots. Provides real-time alerts.	Instant citizen reporting. Identifies hotspots & patterns. Real-time alerts.	Needs better prediction accuracy. Limited data scope. Scalability & privacy concerns.
[8]	Web (PHP, JS, MySQL), Maps API	Generates crime maps from user-reported data. Uses MySQL storage & Google Maps visualization.	Visualizes crime distribution. Easy-to-use interface.	Mapping only (no prediction). Data accuracy is user-dependent. No alerts.
[9]	Java, Python, R, Android, MySQL, Maps API, NLP (BERT, NLTK)	Uses ML (SVM, NB) for classification & NLP (BERT, etc.) for text analysis.	Real-time alerts. Uses text mining for deeper analysis. High classification accuracy.	Scalability & data privacy issues. High model complexity.
[10]	Python (Scikit-learn, Keras), R, Tableau, Big Data (Hadoop, Spark)	A review of various ML/DL models (CNN, RNN, LSTM) for crime prediction & analysis.	Comprehensive review of ML/DL methods. Includes feature engineering.	DL models are computationally expensive. Data sparsity & quality issues.
[11]	Android, Java, PHP, MySQL, Maps API	A standard Android app for crime reporting with a web admin panel.	Simple mobile app for reporting. Includes web admin panel.	Limited features (no prediction). Data security & scalability concerns.
[12]	Android, Java, MySQL, Maps API, GPS, GSM	Uses GPS for location & GSM to send real-time alerts directly to nearby police stations.	Real-time alerts direct to police. Uses GPS for accuracy.	Dependent on GSM network. Security concerns. Reporting only, no analysis.
[13]	Python (Scikit-learn), R, MySQL, Tableau	Focuses on pattern recognition (K-means, Apriori) to find hotspots & frequent patterns.	Identifies hotspots/patterns using unsupervised learning. Good data visualization.	Data quality issues. Limited to pattern recognition, not prediction.
[14]	Android, Java, PHP, MySQL, Maps API, Encryption (AES, RSA)	A secure platform using AES/RSA encryption. Separate interfaces for users & police.	Strong focus on data security & privacy. Real-time reporting.	Encryption adds overhead. Key management is complex. Scalability issues.
[15]	Android, Java, PHP, MySQL, Maps API, GCP, GIS	A cloud-based (GCP) system using GIS for spatial analysis & K-means for hotspots.	Scalable (cloud-based). Advanced GIS analysis. Hotspot detection.	Cloud service costs. Data security on cloud. Requires internet.
[16]	Python (OpenCV, Keras), R, MySQL, Maps API, Android	Uses multimodal data (text, image, video) with DL models (CNN, RNN) for analysis.	Comprehensive analysis using multimodal data. High detection accuracy.	High computational cost. Data fusion is challenging. Privacy concerns.

2.2 Multimodal Data Processing in Crime and Incident Reporting

2.2.1 Existing Portals and Systems in India

The digital crime reporting infrastructure in India has undergone a massive transformation, moving from isolated databases to a centralized, interoperable network.

- Crime and Criminal Tracking Network & Systems (CCTNS): Launched as a mission-mode project under the National e-Governance Plan (NeGP), CCTNS connects over 17,171 police stations across the country. The system aims to computerize policing activities such as registering First Information Reports, investigating crimes, and filing chargesheets. The Digital Police Portal offers citizens services such as e-FIR for reporting lost property and tracking of police complaints in near real-time [17].
- National Cyber Crime Reporting Portal (NCRP): This portal is under the control of the Indian Cyber Crime Coordination Centre (I4C) and is used to coordinate all kinds of cybercrimes. However, one of the key sub-modules is Citizen Financial Cyber Fraud Reporting & Management System (CFCFRMS), which is integrated with the 1930 helpline. This helpline coordinates with over 85 banks for the freezing of transactions during the "golden hour".
- Inter-operable Criminal Justice System (ICJS)
- This platform facilitates the concept of "one data, once entry" through the integration of the five pillars of the criminal justice system: the police, Courts (e-Courts), prisons (e-Prisons), forensics (e-Forensics), and prosecution. This allows for the efficient transfer of FIR details and electronic evidence from police to courts [17].
- Telecom-Centric Fraud Reporting (Sanchar Saathi): The Department of Telecommunications (DoT) launched the Sanchar Saathi app and the Chakshu module, through which citizens can report fraudulent calls, SMS, and WhatsApp messages such as "Digital Arrest Scams," and block the stolen devices using IMEI tracking.

2.3 Limitations in Existing Portals

Despite the success of CCTNS and NCRP, several and technical gaps persist, which your research aims to bridge.

- Text-Centricity and Manual Effort: The majority of existing Indian portals are manual in nature and use form-based entry systems. Victims of crime, being overwhelmed by legal terminology and multi-page digital forms, result in lower reporting of crimes. Currently, there is no mechanism to draft legal reports with raw voice or video clips.
- High Investigative Latency: Since existing systems do not have automated synthesis, policemen must watch every video and listen to every audio file individually to derive the evidence. This has resulted in a huge backlog, particularly in cyber cells that have seen an increased rate of over 24% annually [18].
- Evidence Fragility and Reliability: Although it is possible to upload files, legal usability is difficult to guarantee. In addition, digital evidence is not easily supported in a court of law because of the possibility of tampering through the absence of "evidence linkage," or the tracing of all conclusions obtained by AI to the original data-provider metadata [19].
- The Digital Literacy Gap: Although mobile penetration is high in rural India, digital literacy is reported to be poor, e.g., only 35% among females. Users find it difficult to deal with structured portals. There is a need for a more intuitive interface that can process natural language as well as voice.
- Connectivity Dependency: Existing technologies like NCRP and CCTNS are only available on-line. In "shadow zones," citizens are cut off from law enforcement.

2.4 Proposed Solution: Multimodal AI Architecture and Integration

Your proposed multimodal system addresses the "unstructured data" problem by integrating an end-to-end AI pipeline that automates evidence synthesis.

- Multimodal Fusion & AI Pipeline: The fundamental novelty lies in the AI Processing and Intelligence Layer, where text, audio, images, and video are treated as first-class evidence; whereas in other portals, these are just "stored" as attachments, in your platform through a Multimodal Fusion Engine, these semantic cues are aligned-for example, if a user has vocally described the "red jacket," then is that correct in reality with respect to what comes out visually through an image analysis?
- Large Language Model Integration: The platform utilizes LLMs to parse unedited transcripts and graphical descriptions into highly organized schema-compliant JSON reports. It simplifies the process by detecting types of incidences, parties involved and making chronological summaries of events by automatically detecting the types of incidences, the parties involved in under 5 seconds of extra workload by the police.
- Cross-Modal Validation: There are automated forms of consistency checks used on the platform. An example is that the GPS metadata is compared to the visual indicators of the environment, i.e., street signs or lights, signifying the possibility of discrepancy or deepfake presence.

Resilient Mesh Networking: The platform combines peer-to-peer (P2P) mesh networking [20] to address the connectivity gap that is very common among rural Indian populations. This enables reports to be passed between

devices until a node is reached with the internet connection hence the ability to report during disasters or network failure.

III. METHODOLOGY

3.1 Data Sources and Multimodal Inputs

The proposed system will work based on "real-life" data that is generated by citizens and gathered near the scene of the incident. This may include text descriptions, images, voice recordings, short videos, and location information that can be obtained from the mobile application interface. Each of these will be primary data inputs, aiming to enrich user experience by eliminating reporting barrier as it reflects "natural" usage of smartphones to report an incident. Location and time metadata will automatically be added, as has been seen in past mobile-based crime reporting systems.

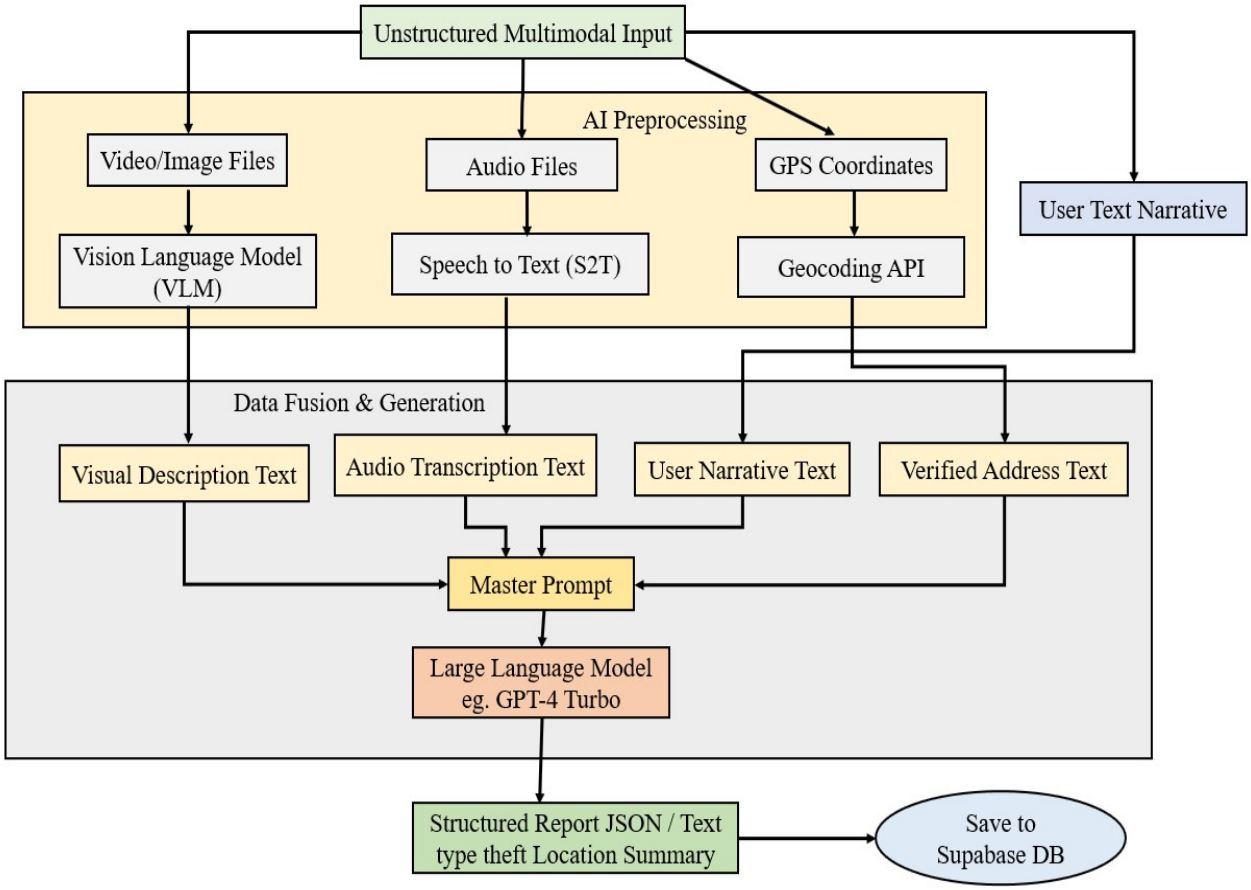

Figure 1. Input Data Flow

The proposed framework treats all modes as first-class sources of information, as opposed to traditional form-based systems that rely only on text modality. Although visual modals provide auxiliary context and corroborating information, audio models are processed using speech-to-text to obtain verbal reports. Cross-validation of information from all modals is enabled by multimodal fusion, hence improving report completeness and trustworthiness. This approach is supported by recent research indicating that while incorporating text, vision, and audio modals in comparison to unimodal systems enhances richer and more useful understanding of the incident.

3.2 Research Design and Approach

The paper is system-centric, design science-based, and aims to create a practical, AI-based reporting system that can be used to address practice limitations in existing crime reporting systems. The research design gives great prominence to problem driven innovation in which system requirements are elicited based on the documented weaknesses in existing crime-reporting system: poor evidence usability, unstructured data, delayed reporting and lack

of anonymity. The paper highlights the importance of combining multimodal data collection, AI-generated interpretation, and structured output creation under a real-life reporting process as opposed to individual algorithmic functionality.

It is approached through an iterative process of prototyping and validation so that the given approach will guarantee technical feasibility and operational applicability. The system architecture integrates voice-to-text audio recordings, language-driven visual proof, and LLMs for lexical report generation into a cohesive pipeline. The assessment methodology comprises scenario testing under control, case-based testing, and qualitative completeness, and usability of reports at the levels of authority-facing dashboards. This is in line with previous work that places more emphasis on end-to-end evaluation than model accuracy in isolation for platforms supporting intelligence generation and digital crime reporting.

3.3 Proposed System Architecture

The proposed system architecture is a layered, service-oriented system, which would allow for unhindered communication between law enforcement, AI processing, and citizens. Scaling, fault tolerating, multimodal intelligence, and law usability remain major concerns with the proposed system, ensuring that any unprocessed incident data collected from the field can be converted into a standard report. Four constituent parts make up the proposed system, including the Citizen Reporting Layer, Connectivity and Data Management Layer, AI Processing and Intelligence Layer, and Authority Analytics and Control Layer [21].

For the real-time interaction in geospatial media, the Citizen Reporting Layer is developed as a cross-platform app using Flutter and integrating it with the Open Street Map. In addition to seeing their current location, users can initiate an incident report and click anywhere to get the latitude and longitude. Various media formats can be accommodated, including structured text input, image uploading, video recording, and audio capture. Recording audio directly can also be done using the program, enabling the audio to be sent for speech-to-text conversion. In order to alleviate security and privacy issues that frequently dissuade individuals from reporting crimes, users can report incidents anonymously or using authentic identities. Using an offline-first approach to app design helps to support data caching and can be sent later to handle emergency situations.

The management of resilience and data transfer is carried out by the Connectivity and Data Management Layer. When internet connectivity is available, reports will be sent over an encrypted API to a centralized backend. A peer-to-peer method of networking is initiated when there is a lack of connectivity between devices, allowing neighboring devices to route reports to a node with internet access. This ensures reports are not interrupted under circumstances such as infrastructural failures, protests, or disasters. A highly controlled database with monitoring capabilities is used to store data, including multimedia data, user inputs, and reports.

The AI processing and intelligence layer makes up the heart of the system. In processing input multimodal data, it first goes through pre-processing by modality-specific models such as speech-to-text models that process audio-based evidence, vision language models that process images and videos, and metadata processing that normalizes spatial and temporal attributes. These results are then fused into a unified form by the multimodal fusion engine, whose purpose is to match evidence between different modalities and address any ambiguity or reliability of results. This is then processed by a Large Language Model that carries out contextual reasoning, fact extraction, and narrative synthesis, ultimately resulting in a structured form of an incident report pre-defined by certain schemas, including type, location, incidence, entities, and finally, evidence inclusion.

Finally, the Authority Analytics and Control Layer is realized as a secure online dashboard for law enforcement. The dashboard employs various graphical user interface elements such as trend graphs, heatmaps, and maps to display live updates of occasions of infraction. Officers get to utilize AI-produced reports, associated multimedia evidence, case assignments, updates on investigation statuses, and past pattern analysis while conducting preventative policing. The inclusion of predictive analytics, language translation, or validation processes is facilitated based on the reporting system itself without altering the underlying reporting process through the modular separation between the two layers, which ensures maintainability.

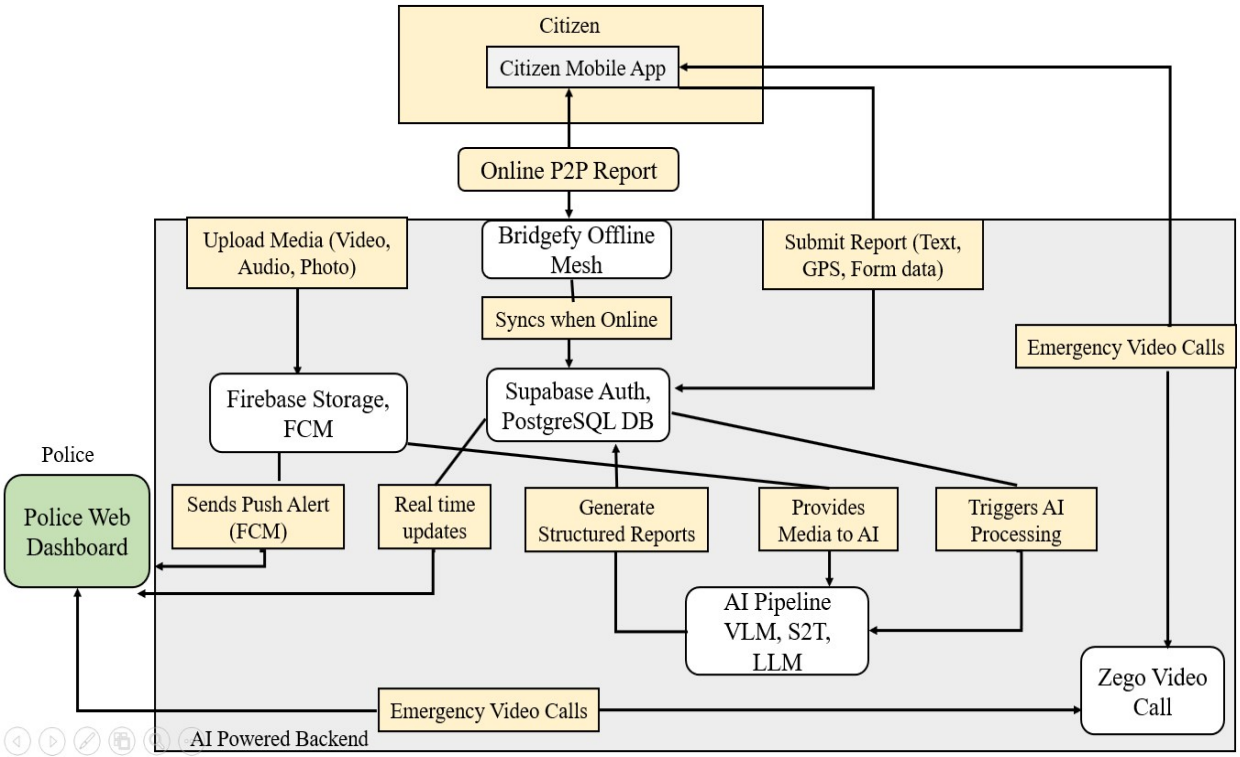

Figure 2. Overview of Proposed Model

3.4 Multimodal AI Processing and Report Generation Pipeline

Through a number of tightly connected processing stages, the multimodal AI pipeline is designed to transform disordered and diversified forms of evidence into well-structured and comprehensible reports in a legal sense. The text, speech, picture, video, and metadata forms of evidence are transferred to the respective preprocessor blocks at the beginning of the processing of a received report. The normalization of text forms guarantees preservation of the intent of the user while discarding noise. The speech-recognition models are utilized for the transcription of audio recordings with the aim of collecting spoken narratives involving both temporal and emotional components. The representative frames of the picture are analyzed through vision and language models with a purpose of extracting contextual, object-related, and environmental evidence elements, such as the scenes of actions and environmental conditions. The geospatial and temporal metadata normalization guarantees the simultaneous standardization of the position and time references.

In order to increase reliability and decrease ambiguity, the system aligns semantic information across modalities using multimodal data fusion after preprocessing. For instance, verbal descriptions from audio transcripts are verified using GPS metadata and cross-validated against visual observations from pictures or videos. The system can find discrepancies, missing information, or supporting evidence thanks to this fusion process, which creates a uniform semantic description of the incident. By addressing the shortcomings of unimodal crime reporting systems noted in earlier research, such cross-modal reasoning improves robustness and completeness.

Such a merged representation is finally passed to a large language model-based report generation module, which carries out contextual reasoning, fact extraction, and narrative synthesis. Here, the large language model is used to automatically generate a structured incident report, adhering to specific schemas. Such a generated report includes aspects such as classification, chronological sequences of events, entities, location, etc., with output format adhering to standard templates. Such automated processes comply with what has been proposed with respect to intelligence reporting with the application of multimodal intelligence reporting, citizen-based reporting of crimes, etc.

3.5 Report Validation and Legal Structuring Strategy

An important factor in the design of the proposed system is ensuring that AI-generated incident reports are trustworthy, feasible, and legally admissible. To achieve this, a validation/legality layer that ensures concepts of consistency, completeness, and evidence integrity is tightly coupled with the incident reporting creation process. Instead of using free-text reporting, which is generated by using a schema-driven organized reporting process, each

component of the incident report, e.g., type of incident, description, entities, evidence, among others, is clearly described. This follows the law enforcement uniform documentation processes.

There are many validation steps throughout the pipeline, too. At the multi-modal fusion step, for example, consistency checks are carried out to ensure that pieces of information obtained from different sources are not inconsistent. For example, pieces of evidence gathered from images that suggest that a crime happened at night are cross-checked against user-provided time stamps, among other metadata. Similarly, location-based metadata are cross-checked against coordinates on maps. Incomplete or misleading reports are filtered to prevent them from passing through unchecked, as has been acknowledged as one of the limitations of citizen-provided crime maps as explored in previous studies on digital crime reporting.

From a legal perspective, the emphasis is on traceability as well as linking evidence. This is because every time a report is produced, its links to the original multimedia inputs will be clearly visible, thereby allowing investigators to trace the source of conclusions based on raw evidence. Such access, audit, as well as immutable timestamps, has been incorporated as part of the backend of such AI-based reports, thereby allowing them to be used as part of law enforcement workflows, thus filling a critical gap between AI-based automation and law enforcement scenarios.

IV. System Implementation and Experimental Setup

4.1 Hardware and Software Specifications

A prototype deployment environment that was reflective of actual operating conditions was used to implement and assess the suggested system. Android smartphones with octa-core processors (⩾2.0 GHz), 4 – 8 GB RAM, 64 – 128 GB internal storage, GPS modules with ⩽5 m location accuracy, and cameras that can capture 12 – 48 MP images and 1080p – 4K video were used to test the mobile application on the client side. In order to ensure enough quality for speech-to-text processing, audio inputs were recorded using integrated microphones with a sample rate of 44.1 kHz. These specs improve accessibility and scalability by reflecting widely available mid-range devices.

A cloud-based server environment with 8–16 vCPUs, 32–64 GB RAM, and scalable object storage for multimedia evidence was used to set up the backend architecture. RESTful APIs were deployed to support data ingestion, authentication and report administration. All the interactions were secured by using HTTPS (TLS 1.2/1.3) encryption. Structured data were kept in relational databases and multimedia files were dealt with in object storage developed to take large amounts of binary data. Audit logging and role-based access control were turned on to ensure integrity of data and compliance.

The AI processing stack uses models for understanding images and videos using computer vision, speech-to-text predictive models with a median latency of 1–2 sec per 30-second audio trial, and a Large Language Model to make organized reports in no more than five seconds for each incident during normal load. Introduced as a web-based application, the authority dashboard enables the effective monitoring and management of the case procedure by providing real-time geospatial visualization and handling hundreds of simultaneous reports on the case with a response time of less than 1 second.

4.2 Core System Components

All the five closely related key elements that constitute the proposed system are tasked with specific functional duties in the end-to-end incident reporting pipeline. By combining all of these factors, one can have intelligent processing, resilient communication, real-time data collection, and actionable display to the authorities.
Citizen Mobile Application is the primary data collection instrument. It was created with Flutter and can take multimodal input such as text (up to several hundred tokens per report), photos (JPEG/PNG up to ~10 MB), audio recordings (WAV/M4A, 44.1 kHz), and short video clips (MP4, 10 – 60 seconds). With an accuracy of ⩽5 m in latitude and longitude, the app incorporates OpenStreetMap for real-time location tracking and map-based incident selection. While optional anonymous reporting increases user engagement, local caching and offline storage guarantee zero data loss during network outages.

Secure data transmission, storage, and orchestration are managed by the Connectivity and Backend Management Component. RESTful APIs are used to send reports across encrypted channels (TLS 1.2/1.3). Multimedia evidence is kept in object storage, and structured incident data is kept in a relational database. With automatic metadata tagging and indexing for quick retrieval, the backend can handle hundreds of reports being ingested concurrently per hour.

Multimodal fusion, vision-language analysis, and speech-to-text transcription are all carried out by the Multimodal AI Processing Component [22]. While visual analysis collects scene context, objects, and activities, audio transcription achieves near-real-time processing (about 1-2 seconds per 30-second clip). Semantic, spatial, and temporal information are aligned by a fusion engine before being sent to the LLM. The LLM-Based Report Generation Component ensures consistency and legal usability by synthesizing fused inputs into structured, schema-compliant reports in three to five seconds per event. Lastly, real-time maps, heatmaps, trend analysis, and case management capabilities are provided to law enforcement by the Authority Analytics and Dashboard Component. Effective operational decision-making and data-driven policing are made possible by its capability for hundreds of concurrent users, sub-second query replies, and secure role-based access.

4.3 Testing and Evaluation Methodology

Evaluation of the system was carried out using multi-stage and scenario-based evaluation criteria. In this regard, three major dimensions were involved in the evaluation process:

i) functionality evaluation of the different components of the system,
ii) performance and latency evaluation of the multimodal AI-based pipeline, and
iii) assessing the robustness of the system under different levels of connectivity.

During the evaluation process, 30+ simulated incident reports were generated to ensure the evaluation and testing of the system in different scenarios.

In other words, from the submission of reports via the mobile application to the visualization of reports via the authority dashboard, functional testing assured us of the accuracy of the "whole workflow." To ensure the ingestion, storage, processing, and retrieval of all information as needed, all individual and collective modality reports, i.e., text, image, audio, video, and GPS, as well as anonymous reporting validation, multimedia linking, and structured report generation, received special consideration. Thus, 100% report delivery, accurate metadata attachment, and output production that adhered to the schema were considered by us as success criteria.

Responsiveness of the system and processing delays were a principal focus of performance evaluation. The synthesis time for generating the reports using the LLM, the inference time per image or video frame for the vision-language model, transcription time for the speech-to-text feature expressed per second per audio clip, and the end-to-end report generation delays were all significant factors. The average end to end processing time was determined under normal load and simultaneous submission conditions (10 to 50 reports/hour). The limits regarding the acceptable response time on the dashboard and map creation delay have been established as less than one second to respond to queries and five seconds to complete a full report.

The connectivity and resilience testing determined the performance of the system when there was a weak or non-existent network. The mesh-based forwarding between peers, creation of reports offline and caching of reports on local devices were also experimented on different devices to ensure that the report was delivered once the connections were reestablished. Besides the quantitative indicators, user feedback in the form of qualitative data was also collected to assess the interpretability, clarity, and reporting convenience on the authority side, which had a supporting role in system efficacy validation.

V. Results and System Performance Analysis

5.1 Multimodal AI Processing Performance

In this section, the computational performance and responsiveness of the proposed AI-driven incident reporting system with multimodal is evaluated. Since processing delay, throughput, and end-to-end responsiveness directly affect real-time usefulness of both civilians and law enforcement officials, performance studies are concerned with these indicators. The experiments involved the use of thirty simulated incident report involving various text, image, audio, video, and GPS metadata mixes to simulate real-world incident reporting scenarios [23].

Audio inputs of approximately 2030 seconds were transcribed with speech-to-text modules with an average latency per clip of 1.21.8 seconds to process them modality-wise. Although 10-15 seconds of short videos required processing time of 1.5-2.2 seconds, including frame sampling and contextual feature extraction picture analysis with

vision-language models required approximately 0.6-1.0 seconds per image. The system led to a low overhead (approximately 0.3 to 0.5 seconds) at the multimodal fusion phase, where semantic, temporal, and spatial information in the modalities are aligned.

The structured report creation phase that used the LLM demonstrated consistent performance and when it first used the schema generated 2.5-3.5 seconds incident report per submission. When the load was normal (10-20 reports per hour), the total end-to-end latency, which was defined as the time interval between the submission of the report in the mobile application and its appearance on the authority dashboard, ranged between 4.8 and 6.5 seconds. The near real-time capability in the deployment of the system was proved through the minor rise in latency at extreme concurrent load (3050 reports/hour) although this was within acceptable operation levels (less than 8 seconds).

Table -2 Structured Report Quality and Usability Analysis

Report Attribute	Required	Average Completion Rate (%)
Incident Type	Yes	100
Location (Lat, Lon)	Yes	100
Timestamp	Yes	96
Incident Description	Yes	94
Audio Transcript Summary	Optional	90
Visual Evidence Reference	Optional	93
Severity Indicators	Yes	88
Involved Entities	Yes	91
Evidence Linking	Yes	95
System Confidence Notes	Optional	89

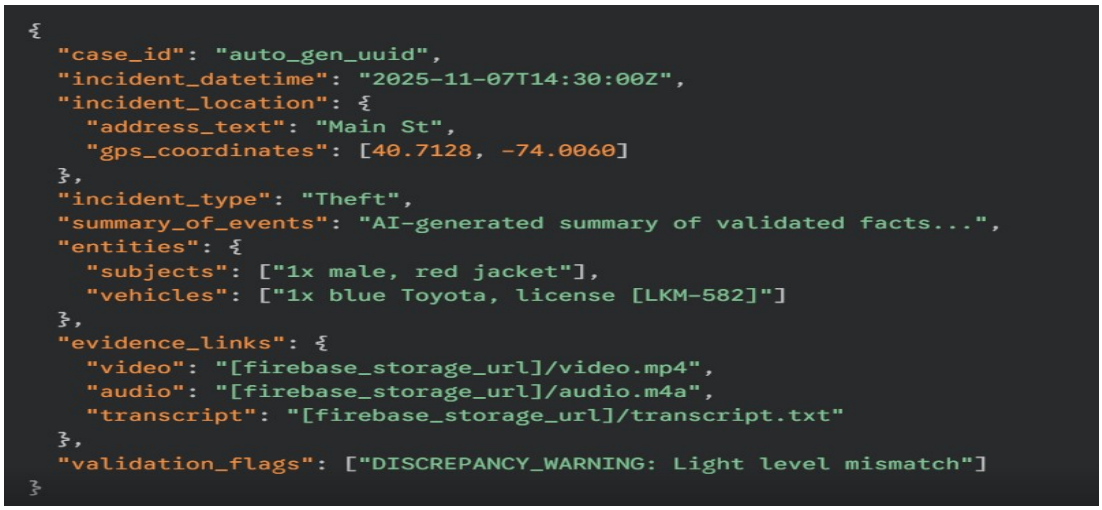

Figure 3. Report Schema

5.2 Case Study Scenarios

Three scenarios in form of case studies were created and implemented to verify the feasibility of the proposed system. These situations represent the reality of workplace and illustrate how the system works in places of offline and anonymous work as well as online. The presence of input modalities, processing flow, system response time and operational output, recorded in each of the case studies provides qualitative and quantitative evidence of efficacy.

Case Study 1: Online Multimodal Reporting in Real Time

In the given case, an individual with live connection to the internet was able to report an incident with the help of GPS metadata, 2 photos, and 150 pieces of text. Once the mobile application identified the current position of the user with a precision of 5 m and less, the report was sent immediately to the backend. The production of structured

reports based on LLCM required approximately 3.0 seconds following image analysis and multimodal fusion that required approximately 2.1 seconds. At a time of about 5.4 seconds after submission, the final report was displayed on the authority dashboard. The authorities have shown close to real time situational awareness and readiness to respond very fast by being able to determine the nature, location and severity of the incidence.

Case Study 2: Anonymous Incident Reporting with Evidence Preservation

The anonymity of a user was evaluated, in this instance, without compromising the integrity of the evidence. The writer preferred to remain anonymous and provided one picture, text and a 25 seconds sound file. The system had the capacity to effectively remove personally identifiable information whilst preserving timestamps, geolocation, evidence references, etc. The whole structured report took about 6.0 seconds to bring up to completion, with the speech to text transcription taking a time of about 1.6 seconds. Although no identity information was provided, authority-side assessment confirmed that the report was completely actionable and that the system had the ability of reducing reporting hesitancy without compromising the legal usability.

Case Study 3: Peer-to-Peer Mesh Networking for Offline Reporting

Something. It was announced by both text and voice messages at the location where the internet was unavailable. This was to determine the effectiveness of the system. The report was dispatched to a device with internet connection through a special type of networking that allows device to communicate to one another directly. The first storage was in the device. Then it was forwarded to the other device. Everything then went on as usual. It took the report to the people in charge between 9 and 12 minutes based on whether the other devices were functioning. The report got there safely. No information was lost or jumbled up. The authority dashboard could view what was given to the incident through the report sent by the authority. This situation confirms the fault tolerance of the system and its suitability in remote area, disaster area or network failure.

VI. Future Directions

Even though the proposed multimodal AI-powered incident detection system is acknowledged for its efficiency and practicality, it has still opened a number of areas where it may be enhanced and further explored. A notable area is making it scalable to a city or state level, where it will be required to manage hundreds, if not thousands, of concurrent incidents, which will require a form of distributed AI system processing. Another area is ensuring its accessibility and inclusion in different communities, especially for communities that speak different languages. This, to a large extent, may be achieved by adding support for multiple languages, including code-mixed languages, particularly for regional languages. Furthermore, to increase the integrity of the evidence for admissibility, advanced methods of validating existing evidence, such as the use of blockchain technology, may be explored. Additionally, intelligent policing methods or interventions may be achieved by utilizing techniques such as predictive analytics combined with spatiotemporal crime predictions and incorporating a form of human-in-the-loop for its reviews, allowing the authority to improve its performance, particularly for incident reports generated by AI.

VII. Conclusion

This study introduced a multimodal AI-based incident reporting system, which overcomes some of the main limitations of traditional methods of reporting crimes. The technology increases the accessibility of reporting and improves the quality of evidence with a single mobile application, including text, image, audio, video, and location-based data. Law enforcement agencies can decrease their workload through a reduction in manual labor using unstructured input data to be processed and converted into a legally acceptable format using Large Language Models and multimodal AI technologies. Enabling offline mesh-based communication and anonymous reporting improves user engagement and system resilience in the event of network outages. Through an authority-facing analytics dashboard, experimental evaluation and case studies show that the suggested solution provides near real-time report generation, enhanced report quality, and actionable intelligence, confirming its efficacy as a scalable and useful digital crime reporting system.

REFERENCES

[1] D. Lal, A. Abidina, N. Garg, and V. Deep, “Advanced Immediate Crime Reporting to Police in India,” *Procedia Computer Science*, vol. 85, pp. 543–549, 2016.

[2] T. Alameri, A. H. Alhilali, N. S. Ali, and J. K. Mezaal, "Crime reporting and police controlling: Mobile and web-based approach for information-sharing in Iraq," *Journal of Intelligent Systems*, vol. 31, pp. 726–738, 2022.

[3] M. Mwiya, J. Phiri, and G. Lyoko, "Public Crime Reporting and Monitoring System Model Using GSM and GIS Technologies: A Case of Zambia Police Service," *Int. J. Computer Science and Mobile Computing*, vol. 4, no. 11, pp. 207–226, 2015.

[4] P. Amudhavalli, S. Rajesh, and N. J. Naseeruddin, "Digital Crime Reporting System by using PHP," *Journal of Applied Information Science*, vol. 10, no. 1, pp. 44–49, 2022.

[5] A. M. Archana, D. S., and K. Saveetha, "Online Crime Reporting System," *International Journal of Advanced Networking & Applications*, pp. 297–300, 2019.

[6] T.-F. Shih, C.-L. Chen, B.-Y. Syu, and Y.-Y. Deng, "A Cloud-Based Crime Reporting System with Identity Protection," *Symmetry*, vol. 11, no. 2, p. 255, 2019.

[7] A. Ahmed, M. Farhan, H. Eesaar, K. T. Chong, and H. Tayara, "From Detection to Action: A Multimodal AI Framework for Traffic Incident Response," *Drones*, vol. 8, no. 12, p. 741, 2024.

[8] A. Wattimena, Harnessing the Power of Generative AI and Multimodal Data Fusion for Smarter Intelligence Reporting, Master's Thesis, 2023.

[9] A. Ahmed et al., "Vision–Language Models for Incident Description and Response Generation," *Drones*, vol. 8, pp. 1–19, 2024.

[10] C.-L. Chen et al., "Identity Protection and Non-Repudiation in Online Crime Reporting Systems," *Symmetry*, vol. 11, pp. 1–29, 2019.

[11] M. Mwiya et al., "Integration of GIS-Based Crime Visualization for Public Safety Applications," *IJCSMC*, vol. 4, no. 11, 2015.

[12] D. Lal et al., "Limitations of Conventional Crime Reporting Systems in India," *Procedia Computer Science*, vol. 85, 2016.

[13] T. Alameri et al., "Challenges in Citizen–Police Information Sharing Platforms," *Journal of Intelligent Systems*, vol. 31, 2022.

[14] P. Amudhavalli et al., "Web-Based Crime Record Management Systems: A Review," *Journal of Applied Information Science*, vol. 10, 2022.

[15] A. Wattimena, "Multimodal Data Fusion for Automated Intelligence Report Generation," 2023.

[16] T.-F. Shih et al., "Cryptographic Mechanisms for Secure and Anonymous Reporting," *Symmetry*, vol. 11, 2019.

[17] A. Ahmed et al., "LLM-Assisted Incident Report Generation from Multimodal Inputs," *Drones*, vol. 8, 2024.

[18] Y. Prajapati, O. P. Suthar, K. Gosai and S. K. Singh, "Smart City Cybersecurity: Leveraging Machine Learning for Advanced Ransomware Detection and Prevention," *2025 International Conference on Pervasive Computational Technologies (ICPCT)*, Greater Noida, India, 2025, pp. 808-813, doi: 10.1109/ICPCT64145.2025.10941048.

[19] M. Mwiya et al., "Mobile Technologies for Crime Detection and Monitoring," *IJCSMC*, vol. 4, 2015.

[20] D. Lal et al., "Immediate Crime Reporting Using Mobile and Relay-Based Communication," *Procedia Computer Science*, vol. 85, 2016.

[21] T. Alameri et al., "Cloud-Based Crime Reporting and Monitoring Platforms," *Journal of Intelligent Systems*, vol. 31, 2022.

[22] National Cybercrime Reporting Portal, Government of India. Available: https://cybercrime.gov.in/ — A centralized online portal for citizens to **report cybercrimes securely and anonymously**, supporting fraud, harassment, and identity theft reporting.

[23] Digital Police Portal, Ministry of Home Affairs, Government of India. Available: https://digitalpolice.gov.in/ — A platform for **online crime complaints and antecedent verification**, offering citizen services and access to national crime record databases.