VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY
An Autonomous Institute Affiliated to University of Mumbai
Department of Computer Engineering

Project Report on

INTEGRATED MULTIMODAL CRIME DETECTION AND PREDICTION SYSTEM

In partial fulfillment of the Fourth Year, Bachelor of Engineering (B.E.) Degree in Computer Engineering at the University of Mumbai Academic Year 2024-25

Submitted by
Chengalva Sai Harikha (D17 - A , Roll no - 05)
Sairaj Deshpande (D17 - A , Roll no - 12)
Anagha Kulkarni (D17 - A , Roll no - 32)
Ketaki Sahasrabudhe (D17 - A , Roll no - 53)

Project Mentor
Dr. Mrs. Sujata Khedkar

(2024-25)

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY
An Autonomous Institute Affiliated to University of Mumbai
Department of Computer Engineering

Certificate

This is to certify that ***Chengalva Sai Harikha(5),Sairaj Deshpande(12), Anagha Kulkarni(32) and Ketaki Sahasrabudhe(53)*** of Fourth Year Computer Engineering studying under the University of Mumbai have satisfactorily completed the project on "***INTEGRATED MULTIMODAL CRIME DETECTION AND PREDICTION SYSTEM***" as a part of their coursework of PROJECT-II for Semester-VIII under the guidance of their mentor ***Dr. Mrs. Sujata Khedkar*** in the year 2024-25.

This thesis/dissertation/project report entitled (***INTEGRATED MULTIMODAL CRIME DETECTION AND PREDICTION SYSTEM***) by ***Chengalva Sai Harikha(5),Sairaj Deshpande(12), Anagha Kulkarni(32) and Ketaki Sahasrabudhe(53)*** is approved for the degree of BE Computer Engineering.

Programme Outcomes	Grade
PO1,PO2,PO3,PO4,PO5,PO6,PO7, PO8, PO9, PO10, PO11, PO12 PSO1, PSO2	

Date:

Project Guide:

Project Report Approval
For
B. E (Computer Engineering)

This project report entitled ***INTEGRATED MULTIMODAL CRIME DETECTION AND PREDICTION SYSTEM*** by ***Chengalva Sai Harikha(5),Sairaj Deshpande(12), Anagha Kulkarni(32) and Ketaki Sahasrabudhe(53)*** is approved for the degree ofBE Computer Engineering.

Internal Examiner

External Examiner

Head of the Department

Principal

Date:
Place:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

---	---
---	---
(Chengalva Sai Harikha - 05)	(Sairaj Deshpande - 12)
---	---
---	---
(Anagha Kulkarni - 32)	(Ketaki Sahasrabudhe - 53)

Date:

ACKNOWLEDGEMENT

We are thankful to our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to Associate Professor (**Dr. Mrs. Sujata Khedkar**) for her kind help and valuable advice during the development of project synopsis and for her guidance and suggestions.

We are deeply indebted to Head of the Computer Department **Dr.(Mrs.) Nupur Giri** and our Principal **Dr. (Mrs.) J.M. Nair ,** for giving us this valuable opportunity to do this project.

We express our hearty thanks to them for their assistance without which it would have been difficult in finishing this project synopsis and project review successfully.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support and selfless help throughout the project work. It is great pleasure to acknowledge the help and suggestion, which we received from the Department of Computer Engineering.

We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement at several times.

Computer Engineering Department
COURSE OUTCOMES FOR B.E PROJECT

Learners will be to,

Course Outcome	Description of the Course Outcome
CO 1	Able to apply the relevant engineering concepts, knowledge and skills towards the project.
CO2	Able to identify, formulate and interpret the various relevant research papers and to determine the problem.
CO 3	Able to apply the engineering concepts towards designing solutions for the problem.
CO 4	Able to interpret the data and datasets to be utilized.
CO 5	Able to create, select and apply appropriate technologies, techniques, resources and tools for the project.
CO 6	Able to apply ethical, professional policies and principles towards societal, environmental, safety and cultural benefit.
CO 7	Able to function effectively as an individual, and as a member of a team, allocating roles with clear lines of responsibility and accountability.
CO 8	Able to write effective reports, design documents and make effective presentations.
CO 9	Able to apply engineering and management principles to the project as a team member.
CO 10	Able to apply the project domain knowledge to sharpen one's competency.
CO 11	Able to develop professional, presentational, balanced and structured approach towards project development.
CO 12	Able to adopt skills, languages, environment and platforms for creating innovative solutions for the project.

Index

Chapter 5: Implementation of the Proposed System

5.1. Methodology employed for development

5.2 Algorithms and flowcharts for the respective modules developed

5.3 Datasets source and utilization

Chapter 6: Testing of the Proposed System

6.1 . Introduction to testing

6.2. Types of tests Considered

6.3 Various test case scenarios considered

6.4. Inference drawn from the test cases

Chapter 7: Results and Discussion

7.1. Screenshots of User Interface (UI) for the respective module

7.2. Performance Evaluation measures

7.3. Input Parameters / Features considered

7.4. Graphical and statistical output

7.5. Comparison of results with existing systems

7.6. Inference drawn

Chapter 8: Conclusion

8.1 Limitations

8.2 Conclusion

8.3 Future Scope

References

Appendix

1. Paper I & II Details

a. Paper published
b. Certificate of publication
c. Plagiarism report
d. Project review sheet

2. Competition certificates

Abstract

This project addresses the limitations of traditional crime detection methods by introducing an integrated crime detection and prediction system that utilizes Large Language Models (LLMs) and AI Agents alongside video, audio, and image data. The system combines advanced machine learning techniques with real-time multimedia analysis and historical crime data to create a comprehensive solution for crime prevention and law enforcement. By analyzing textual data from various sources, such as social media, the system improves the accuracy of crime detection and classification while enhancing contextual understanding. It allows law enforcement agencies to upload images of suspects, detect key features such as age, gender, height, weight, and BMI, and also summarize the images, identify weapons, and generate detailed crime reports. Key features of the system include the classification and summarization of crime reports using LLMs, the extraction and analysis of FIR documents via OCR, and real-time face recognition for suspect identification. Additionally, the system includes an AI agent-powered video summarizer that detects violent activity in crime-related videos and responds to user queries. This integrated approach significantly enhances public safety by enabling the system to detect, analyze, and alert users about threats from various forms of media.

Chapter I: Introduction

The Integrated Multimodal Crime Detection and Prediction System aims to combine multiple data sources and advanced machine learning techniques to enhance the efficiency of crime prevention and response. By utilizing modalities such as text, image, and audio data, the system enables comprehensive analysis, real-time detection, and accurate predictions of criminal activities. This project seeks to address current limitations in traditional crime detection methods, offering an innovative approach that leverages technology for improved safety and security.

1.1 Introduction

In today's world, law enforcement agencies face significant challenges in managing the vast and diverse data generated from various sources, such as surveillance cameras, social media, and public records. The sheer volume and complexity of this information can overwhelm traditional crime detection methods, often leading to delayed responses or missed threats.

To address these challenges, our project - Integrated Multimodal Crime Detection and Prediction System offers a comprehensive solution designed to enhance crime prevention and detection capabilities. This innovative system integrates data from multiple modalities including video surveillance, audio recordings, text descriptions, and facial recognition—to provide a holistic view of potential criminal activities.

By consolidating and analysing these varied data inputs, the system aims to deliver a more accurate and insightful analysis of crime-related events. It not only detects incidents where criminal activities may have occurred but also offers predictive insights, enabling proactive measures to prevent potential crimes. This multimodal approach not only improves the accuracy of crime detection but also enhances the system's ability to classify and predict criminal behaviour, offering valuable insights into emerging threats and potential risks.
Through this approach, the system empowers users to make informed decisions and take timely actions, ultimately contributing to safer communities and more effective law enforcement.

Through this innovative approach, the Integrated Multimodal Crime Detection and Prediction System represents a significant leap forward in the field of law enforcement, providing a powerful and effective tool for safeguarding communities and ensuring a more secure future.
In addition to integrating diverse data sources, the Integrated Multimodal Crime Detection and Prediction System utilizes advanced machine learning algorithms and artificial intelligence to ensure seamless and

efficient processing of large datasets. By combining deep learning models with traditional crime analytics, the system automates the detection and classification of suspicious activities, reducing human error and ensuring real-time responsiveness.

The system leverages real-time data streaming from multiple sources, such as surveillance cameras and social media platforms, allowing it to dynamically adapt to the evolving security landscape. With the increasing use of public records and social media by criminals to organize or communicate covertly, the system's natural language processing (NLP) algorithms can parse and analyze text for keywords and patterns that indicate potential criminal intent, ensuring that authorities are alerted to potential threats even before they materialize.

Moreover, the predictive capabilities of the system are grounded in sophisticated time-series forecasting and pattern recognition models that analyze historical crime data to predict where and when crimes are likely to occur. By identifying hotspots and emerging trends, law enforcement agencies can allocate resources more efficiently and deploy officers in high-risk areas, reducing the likelihood of incidents.

1.2 Motivation

Modern cities face increasingly complex crime patterns that cannot be effectively addressed by traditional law enforcement techniques. As crime evolves, especially with the rise of cybercrime and organized criminal networks, existing systems struggle to keep up. There is a pressing need for tools that can analyze large datasets and identify crime trends in real-time. This project is motivated by the potential of LLMs to provide such solutions by processing diverse data sources and understanding the complex nature of criminal behavior.The ability to anticipate and prevent crimes in increasingly complex urban environments is critical for maintaining public safety and improving the quality of life in cities. Also, current crime prediction systems often suffer from a lack of precision, resulting in either an overestimation or underestimation of potential threats. This can lead to either unnecessary allocation of resources or failure to prevent actual crimes. By leveraging the powerful language understanding capabilities of LLMs, this project aims to improve predictive accuracy, reducing the number of false positives while identifying high-risk areas and individuals more effectively. Biases in traditional crime detection systems are a significant issue, often resulting in the disproportionate targeting of certain communities. These biases are typically embedded in the historical data used to train existing systems. The motivation behind this project is to develop an AI-based system that minimizes these biases by using advanced techniques to ensure fairness in predictions. A fine-tuned LLM can analyze data in more nuanced ways, reducing reliance on biased patterns.

1.3 Problem Definition

The project aims to develop an Integrated Multimodal Crime Detection and Prediction System that addresses the limitations of traditional crime detection methods by consolidating diverse data sources such as video, audio, and facial expressions. This approach seeks to improve the accuracy of crime classification and forecasting, enabling more reliable detection of criminal activities and providing predictive insights to prevent potential incidents. The system takes various types of data such as audio, video, text, images as inputs which may or may not contain data related to crime. After the system has been feeded with the input, the system tries to find out and detects if the input has any type of incident where crime has taken place, thus alerting the user about it and making the user take appropriate actions against it.

To improve accuracy and robustness, the system uses a decision fusion layer that consolidates the outputs of individual modalities, enabling a more holistic crime detection mechanism. Additionally, the predictive component of the system leverages time-series analysis and forecasting models to anticipate future incidents based on historical patterns, trends, and real-time data streams.

The system is designed to function in real-time, providing alerts and recommendations to security personnel or law enforcement. These alerts are generated when the system detects patterns consistent with criminal behavior or escalating situations, allowing for timely intervention. The system can be deployed across various environments, including urban surveillance systems, public transport, retail spaces, and critical infrastructure, contributing to proactive crime prevention efforts.

Furthermore, privacy and ethical considerations are embedded within the system, ensuring that data is processed securely and that surveillance activities comply with legal regulations and human rights frameworks. This scalable and adaptive system aims to revolutionize public safety by leveraging multimodal data for more precise crime detection and prevention.

1.4 Existing System

1. Predictive Policing Systems: Predictive policing uses data analytics to forecast potential criminal activity in certain areas. These systems rely on historical crime data, real-time surveillance, and statistical models to predict where crimes are most likely to occur. The two major types of predictive policing models are **location-based** (predicting where crimes will happen) and **person-based** (predicting individuals who might commit crimes).

- **Example**: The **PredPol** system in the U.S. analyzes past crime reports to predict potential future crime hotspots. It uses historical data to identify patterns of criminal behavior in specific locations, helping law enforcement allocate resources more effectively.

2. Crime Mapping Systems: Crime mapping systems visualize criminal activity across a geographic area. By plotting crimes on maps, law enforcement agencies can identify crime hotspots and analyze trends over time. These systems often work in conjunction with other data tools to provide insights into crime patterns.

- **Example**: The **CompStat** system used by the New York Police Department (NYPD) is an example of an early crime mapping tool. It uses data-driven analysis to track and reduce crime by mapping crime data and analyzing trends..

3. Surveillance and Monitoring Systems: Surveillance systems involve using cameras, drones, and sensors to monitor real-time public activities and detect suspicious behavior. In some cities, AI-based video analytics systems can automatically detect and alert authorities about unusual activity.

- **Example**: **ShotSpotter**, used in several U.S. cities, is a system that detects and locates gunshots using acoustic sensors. The data is fed into a central system, where real-time alerts are sent to law enforcement to respond to incidents quickly.

1.5 Lacuna of the existing systems

Despite the advancements in crime detection and prevention technologies, existing systems still face several limitations that hinder their effectiveness in real-world scenarios. Traditional crime detection methods often rely heavily on single-modal data sources, such as text-based reports or isolated video feeds, which may provide only a partial view of the situation. This fragmented approach can lead to incomplete analysis, missed connections, and inaccurate conclusions, ultimately reducing the system's ability to detect and prevent criminal activities effectively.

One significant limitation is the inability of current systems to integrate and process diverse data types simultaneously. For instance, while some systems can analyse video footage for suspicious behaviour, they may lack the capability to correlate this information with audio cues, text descriptions, or other contextual data. This siloed approach limits the system's ability to generate a comprehensive understanding of potential threats, leading to delayed responses or incorrect assessments.
Another challenge is the high volume of data generated by modern surveillance and monitoring tools. Law enforcement agencies are often overwhelmed by the sheer quantity of information, making it difficult to identify relevant patterns or trends in a timely manner.

Existing systems may struggle to filter out noise or irrelevant data, resulting in false positives or negatives that undermine the accuracy and reliability of crime detection efforts.

Furthermore, many current systems lack the advanced predictive capabilities needed to anticipate and prevent crimes before they occur. While they may be effective at detecting incidents after they happen, their ability to foresee potential threats and enable proactive interventions is limited. This gap in predictive analytics means that law enforcement agencies often find themselves reacting to crimes rather than preventing them, reducing their overall effectiveness in maintaining public safety.

The scalability and adaptability of existing crime detection systems also pose significant challenges. Many systems are designed for specific use cases or environments, making it difficult to adapt them to new or evolving threats. As criminal behaviour becomes increasingly sophisticated and diverse, the inability of these systems to learn from new data or integrate with emerging technologies further exacerbates the research gap.Finally, issues related to data privacy, bias, and ethical concerns are prevalent in current systems. The use of AI and machine learning in crime detection can inadvertently reinforce biases present in training data, leading to discriminatory outcomes. Moreover, the collection and processing of large volumes of personal data raise concerns about privacy and the potential misuse of information, which must be addressed to maintain public trust.

These limitations underscore the need for a more integrated, multimodal approach to crime detection and prediction—one that can process and analyse diverse data sources holistically, offer advanced predictive insights, and adapt to the changing landscape of criminal activities. The Integrated Multimodal Crime Detection and Prediction System aims to fill these gaps by providing a comprehensive solution that enhances the accuracy, reliability, and proactivity of crime prevention efforts.

1.6 Relevance of the Project

The United Nations estimates that 68% of the world's population will live in urban areas by 2050, increasing the need for smart crime prevention solutions that can scale with city growth.As cities grow in size and population, the complexity of criminal activities has also increased. Traditional crime detection methods are often reactive and limited in scope, leaving law enforcement agencies struggling to keep up with modern criminal trends. This project is relevant because it offers a proactive approach to crime detection by utilizing LLMs that can analyze vast amounts of real-time data, uncover hidden crime patterns, and predict potential criminal activities before they occur. In many urban environments, crime rates continue to pose significant threats to public safety and quality of life. While traditional law enforcement methods focus on addressing crimes after they occur, this project is highly relevant because it shifts the focus to predictive crime prevention. By fine-tuning an LLM to detect patterns in crime data, social media, and environmental factors, law enforcement can take preventive actions, potentially reducing crime rates and enhancing public safety.

Chapter II: Literature Survey

The literature survey reviews key research papers that focus on crime detection and prediction systems. It provides a brief overview of each paper's contributions and outlines the insights gained from them. By examining various methodologies and technologies used in previous studies, this survey identifies the strengths and limitations of existing systems, which informs the direction of the proposed integrated multimodal approach.

A. Brief Overview of Literature Survey

The literature survey aims to explore the current state of research and development in the fields of crime detection and prediction, particularly focusing on the application of Large Language Models (LLMs) and artificial intelligence. This overview synthesizes key findings, methodologies, and trends observed in recent studies, highlighting the gaps and opportunities for improvement.

Paper [1]explores using large language models (LLMs) for zero-shot crime detection and classification from textual descriptions of surveillance videos.While Paper [2] studies LLM models like GPT-3 and GPT-4 that can surpass traditional machine learning models, such as random forests, in crime classification and prediction using historical data.Paper[5] assesses LLMs for content moderation, finding GPT-3.5 effective in rule-based moderation and showing LLMs outperform current toxicity detectors. However, larger models offer only slight improvements in toxicity detection. Following section includes the detailed description of all the research papers referred.

2.1 Research Papers Referred

Sr.No	Title	Dataset	Models used	Inference
1.	Garbage in, garbage out: Zero-shot detection of crime using Large Language Models	UCF Crime dataset	LLM (GPT-4), For automatic image to text-> 1) Generative Image-to-text Transformer (GIT), LLaVA Descriptions,YOLO-v8 + ByteTrack	This paper explores using large language models (LLMs) for zero-shot crime detection and classification from textual descriptions of surveillance videos. While LLMs achieve state-of-the-art performance when provided with high-quality, manually created descriptions, current automated video-to-text methods produce insufficiently accurate descriptions, leading to poor reasoning results.
2.	A Framework for LLM-Assisted Smart Policing System	Crime data from San Francisco (SF) and Los Angeles (LA)	Compare the ability of the LLMs and ML models(random forest,XGBoost models) in classification and prediction tasks.prompting and fine-tuning methods were used to interact with LLMs, such as GPT models and BART, to analyse their abilities in crime classification and prediction tasks	This study shows that LLMs like GPT-3 and GPT-4 can surpass traditional machine learning models, such as random forests, in crime classification and prediction using historical data. The researchers preprocess crime data to address issues like missing values, and employ prompt engineering to convert structured data into a natural language format for LLMs.
3.	EFFICACY OF UTILIZING LARGE LANGUAGE MODELS TO DETECT PUBLIC THREAT POSTED ONLINE	Extracting 500 post titles from the renowned online platform "DC Inside"1 , specifically from the "실시간 베스트 갤러리" (Real-time Best Gallery). A specialised scraping tool was used to	1) OpenAI’s gpt-3.5-turbo-1106 and gpt-4, as well as PaLM API’s chat-bison. 2)chi-square goodness of fit test at a general significance level of 0.05, was conducted to determine the suitability of employing these LLMs,	This paper evaluates the effectiveness of large language models (LLMs) in detecting public threats online. LLMs like GPT-3.5, GPT-4, and PaLM were prompted to classify posts as "threat" or "safe," with statistical analysis showing all models achieved strong accuracy, passing chi-square tests for both categories.

		exclude any posts containing public threat content from this dataset [30].		
4.	Experimental Analysis of Large Language Models in Crime Classification and Prediction	datasets from San Francisco and Los Angeles	BART, GPT-3, and GPT-4	This paper explores the potential of LLMs like BART, GPT-3, and GPT-4 in smart policing, particularly for crime analysis and predictive policing. While LLMs have been used in various fields, their application in crime classification is underexplored. Using zero-shot, few-shot prompting, and fine-tuning, the study evaluates these models, showing that GPT models outperform traditional ML techniques in most crime classification scenarios.
5.	Watch Your Language: Investigating Content Moderation with Large Language Models		BERT (Bidirectional Encoder Representations from Transformers) T5 (Text-To-Text Transfer Transformer)	This study assesses LLMs for content moderation, finding GPT-3.5 effective in rule-based moderation and showing LLMs outperform current toxicity detectors. However, larger models offer only slight improvements in toxicity detection. Further research in LLMs for moderation is recommended.
6.	Instruction Tuning for Large Language Models: A Survey	1) Super-Natural Instructions: A multilingual dataset with 1,616 NLP tasks and 5 million instances, including definitions and examples . 2) MIMIC-IT: A dataset for	LLaMa, ChatGPT, GPT-4, MultiModal-GPT	The findings indicate that instruction tuning enhances model accuracy and performance across various domains, including dialogue and information extraction. The paper also acknowledges potential pitfalls, calling for further research to improve these methods and better align models with user expectations.

		multimodal instruction-response pairs . 3) Community Q&A Datasets: Includes data from platforms like Stack Exchange and wikiHow, along with manually created examples .		
7.	ChatGPT as a Copilot for Investigating Digital Evidence	-	ChatGPT	The paper discusses the application of ChatGPT in the context of digital evidence investigations, focusing on how it can assist in formulating structured queries, summarising information, and analysing search results based on natural language input from investigators.
8.	GPT-4 Technical Report	-	GPT4	Training: Utilises supervised fine-tuning and reinforcement learning for improved responses . Safety: Implements strategies to reduce harmful outputs . Performance: Outperforms previous models in truthfulness and NLP tasks . Multimodal: Capable of processing text and images . Overall, GPT-4 enhances AI capabilities and safety.
9.	A BERT-Based model: Improving Crime NewsDocuments Classification through AdoptingPre-trained Language Models	Malaysian National News Agency (BERNAMA) and was manually labelled by crime investigation experts into 12	BERT	The paper presents a BERT-based model for classifying crime news documents, addressing challenges such as low efficiency and limited high-quality labelled data. The approach enhances the speed of updating crime statistics and facilitates statistical analysis of crime trends, ultimately contributing to improved public

		categories, including a non-crime class.		safety and crime prevention strategies.
10.	An LLM-driven Approach to Gain Cybercrime Insights with Evidence Networks	The dataset used in the study consists of digital evidence extracted from an Android 10 mobile phone. Specifically, the researchers reconstructed the Forensic Intelligence Graph (FIG) using data from three folders containing three popular Android apps: Phone, Facebook Messenger, and Snapchat	gpt-4-turbo as the supporting Large Language Model (LLM) for their approach to constructing Forensic Intelligence Graphs (FIGs) from digital forensic evidence .	an automated approach for gaining criminal insights with digital evidence networks. This thrust will harness Large Language Models (LLMs) to learn patterns and relationships within forensic artefacts, automatically constructing Forensic Intelligence Graphs (FIGs). These FIGs will graphically represent evidence entities and their interrelations as extracted from mobile devices, while also providing an intelligence-driven approach to the analysis of forensic data. Our preliminary empirical study indicates that the LLM-reconstructed FIG can reveal all suspects' scenarios, achieving 91.67% coverage of evidence entities and 93.75% coverage of evidence relationships for a given Android device.
11.	The Use of Large Language Models (LLM) for Cyber Threat Intelligence (CTI) in Cybercrime Forums	-	OpenAI GPT-3.5-turbo-16k-0613 model for extracting and summarising cyber threat intelligence (CTI) information from cybercrime forums.	the use of Large Language Models (LLMs) for analysing cyber threat intelligence (CTI) data from cybercrime forums. The study evaluates the accuracy of an LLM based on OpenAI's GPT-3.5-turbo model to extract CTI information from 500 conversations across three forums. The LLM achieved an impressive average accuracy of 98%, although the study also identifies areas for improvement, such as distinguishing between stories and events.

12.	MACAW-LLM: MULTI-MODAL LANGUAGE MODELLING WITH IMAGE, AUDIO, VIDEO, AND TEXT INTEGRATION	Text instruction dataset: For textual instruction-tuning, we make use of the Alpaca instruction dataset, comprising approximately 52,000 instruction-response. • Image instruction dataset: To create an image instruction dataset, we curate around 69K instruction-response pairs by generating them from COCO image captions using GPT-3.5-TURBO as described. • Video instruction data: We generate approximately 50K video instruction-response examples by utilizing the video captions from the Charades and	Multimodal: CLIP-ViT-B/16 (Images), WHISPER (Audio), WHISPER-BASE, LLAMA-7B(Text), Models: GPT-4	Although instruction-tuned large language models (LLMs) have exhibited remarkable capabilities across various NLP tasks, their effectiveness on other data modalities beyond text has not been fully studied. In this work, we propose MACAW-LLM, a novel multi-modal LLM that seamlessly integrates visual, audio, and textual information. MACAW-LLM consists of three main components: a modality module for encoding multi-modal data, a cognitive module for harnessing pretrained LLMs, and an alignment module for harmonising diverse representations. Our novel alignment module seamlessly bridges multi-modal features to textual features.

		AVSD.		
13.	Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding	Video-LLaMA 1	Video-LLaMA1	Video-LLaMA1: a multi-modal framework that empowers Large Language Models (LLMs) with the capability of understanding both visual and auditory content in the video.
14.	VLM-Eval: A General Evaluation on Video Large Language Models	Video-LLaVA	Video, Text, Images based LLMs.(Video LLama, ImageBind, GPT-4)	This paper presents a unified evaluation of video Large Language Models (LLMs) across tasks like captioning, Q&A, retrieval, and action recognition. It highlights how GPT-based evaluation can rival human assessment of response quality. The proposed baseline, Video LLaVA, uses a single linear projection and outperforms existing models. Additionally, video LLMs demonstrate strong recognition and reasoning abilities in driving scenarios with minimal fine-tuning.
15.	PG-Video-LLaVA: Pixel Grounding Large Video-Language	PG-Video-LLaVA	Video Instruct 100K dataset comprising 100K video instructions derived from ActivityNet-200	The paper presents PG-Video-LLaVA, a model that enhances video understanding by integrating pixel-level grounding and audio cues. It performs well in video-based tasks and introduces new benchmarks for object grounding and conversation.

16.	A Comprehensive Review of Multimodal Large Language Models: Performance and Challenges Across Different Tasks		Gemini, GPT-4V, NExT-GPT, Video-LLaVA	Multimodal Large Language Models (MLLMs) stand at the forefront of artificial intelligence (AI) systems. Designed to integrate diverse data types—including text, images, videos, audio, and physiological sequences—MLLMs address the complexities of real-world applications far beyond the capabilities of single-modality systems. In this paper a comparative analysis is provided of the focus of different MLLMs in the tasks, and provide insights into the shortcomings of current MLLMs, and suggest potential directions for future research.

2.2. Inference drawn

Enhanced Accuracy Through Multimodal Integration

The system leverages a combination of multiple data sources—such as surveillance footage, social media activity, geographic data, and historical crime records—leading to more accurate predictions and real-time crime detection. By integrating these diverse modalities, the system improves its ability to detect anomalies, suspicious behavior, and potential crime events that might be missed if only one type of data were considered.

Early Detection of Criminal Activities

The predictive modeling capabilities of the system enable law enforcement agencies to identify crime hotspots and potential criminal behavior before incidents occur. This proactive approach helps in early intervention, reducing the likelihood of criminal activities and allowing for better resource allocation to high-risk areas.

Real-time Decision-making

The system's ability to process data from multiple modalities in real time offers substantial improvements in decision-making speed and accuracy. Law enforcement can respond swiftly to emerging threats, deploying resources in real time to areas where the system predicts potential crime or detects suspicious activity.

Adaptability and Scalability

One of the major inferences drawn from the study is the system's adaptability to different urban environments and its scalability to larger geographic areas. Whether deployed in small urban centers or large metropolitan areas, the system can scale accordingly, adapting to varying types and volumes of data.

Minimization of Human Bias

By relying on machine learning models and statistical analysis, the system helps in reducing the bias that may occur in manual crime detection processes. It provides an evidence-based approach to crime prediction and detection, focusing purely on data-driven insights, thereby enhancing fairness and objectivity in law enforcement.

Chapter III: Requirement Gathering for the Proposed System

This chapter outlines the essential requirements for the development of the Integrated Multimodal Crime Detection and Prediction System. It covers the functional and non-functional needs that the system must fulfill to operate effectively, along with the hardware, software, and technologies required for implementation. The process of requirement gathering ensures that the system meets both the technical specifications and user expectations.

3.1 Introduction to requirement gathering

In this phase of the project, our objective is to gather comprehensive information from various sources, ensuring we cover all critical aspects of crime detection, prediction systems, and the use of Large Language Models (LLMs). The goal is to build a strong foundation for the system's development by synthesizing knowledge from academic research, industry insights, technical discussions, and expert recommendations.

1. **Current State of Crime Prediction Systems:**

   A detailed review of existing research papers, reports, and technical literature on crime detection and prediction models will provide a clear understanding of the current state-of-the-art techniques. This includes insights into existing machine learning-based systems, statistical models, and advancements in natural language processing (NLP) as they relate to crime detection. A focus on the strengths and weaknesses of these systems will help identify gaps that our project can address.

2. **Technological Innovations in AI and NLP:**

   Analyzing industry publications, technical blogs, and AI research journals will offer insights into the latest innovations in large language models, deep learning techniques, and advancements in natural language understanding. Monitoring trends in AI, such as transformer models and real-time data processing, will provide valuable perspectives on how to enhance the predictive accuracy and functionality of the crime detection system.

3. **Law Enforcement Practices and Public Safety Requirements:**

   Engaging with law enforcement professionals, criminal justice experts, and public safety organizations will offer practical insights into the real-world application of crime prediction systems. Through workshops, interviews, and surveys, we can gather information about their needs, operational challenges, and expectations from AI-based systems. This interaction will also help in aligning the project’s objectives with law enforcement and community safety goals.

3.2 Functional Requirements

1. **Data Ingestion and Preprocessing:**

   The system must have the ability to gather and preprocess data from various sources such as crime records, social media feeds, environmental sensors, and public reports. This includes cleaning, normalizing, and structuring the data for analysis.

2. **Crime Pattern Recognition:**

   The system should be capable of analyzing historical crime data and recognizing patterns, such as common times and locations for certain types of crimes. It should leverage LLMs to identify trends in textual data sources (e.g., news reports, social media).

3. **Real-Time Crime Prediction:**

   The system should provide real-time crime predictions based on current data inputs, allowing law enforcement to proactively respond to potential threats. It must be able to update its predictions dynamically as new data becomes available.

4. **Integration with Law Enforcement Systems:**

   The system should seamlessly integrate with existing law enforcement platforms (e.g., databases, dashboards, Geographic Information Systems) to enable easy access to predictive insights and crime analytics.

3.3 NonFunctional Requirements

1. **Performance and Response Time:** The system must process large volumes of data and generate crime predictions within a minimal response time, ideally in real-time. The latency for generating insights or predictions from the data should be less than 2 seconds to ensure timely decision-making for law enforcement.

2. **Scalability:** The system should be highly scalable, supporting increased data input as more cities or regions adopt the system. It must efficiently handle growing datasets from multiple data sources, including video surveillance, social media, crime reports, and IoT devices, without performance degradation.

3. **Availability and Reliability:**The system must be available 24/7, with minimal downtime to ensure uninterrupted service. It should have a high uptime of at least 99.9% and be capable of recovering from system failures swiftly to avoid interruptions in crime prediction services.

4. **Security:** The system must follow strict security protocols to safeguard sensitive crime and personal data. This includes encrypted data storage and transfer, secure user authentication, access control, and regular vulnerability assessments to prevent unauthorized access or data breaches.

5. **Maintainability:** The system should be designed for ease of maintenance, ensuring that updates, bug fixes, and improvements can be implemented with minimal disruption to the service. It should include clear documentation for system administrators and developers to manage ongoing updates, including changes in the LLM models or data sources.

3.4. Hardware, Software , Technology and tools utilized

1. Hardware Requirements:

1. GPUs (Graphics Processing Units):

   To fine-tune and deploy LLMs efficiently, we used powerful GPU such as NVIDIA. This GPU provides the necessary computational power for training and inference of deep learning models.

2. **Storage:**

   Large-scale data storage solutions, like **SSD** or **NAS (Network-Attached Storage)**, were utilized to store vast amounts of crime records, social media data, and model checkpoints.

2. Software Requirements:

1. Operating System:

   The servers running the LLM were configured with a stable Linux-based OS like **Ubuntu**, **CentOS**, or **Red Hat** to support GPU acceleration and large-scale data handling efficiently.

2. CUDA and cuDNN Libraries:

   For optimal GPU performance, **CUDA** (Compute Unified Device Architecture) and **cuDNN** (CUDA Deep Neural Network Library) libraries were installed. These libraries enable efficient GPU acceleration for model training and inference.

Chapter IV: Proposed Design

This chapter presents the architectural design for our Integrated Multimodal Crime Detection and Prediction System. It details the system components, their interactions, and the overall workflow that enables efficient processing of multimodal data, including text, audio, video, and images. It includes the user interface design explanation, backend processing modules, and the data flow architecture, ensuring the system aligns with both performance goals and usability standards.

4.1 Block diagram of the system

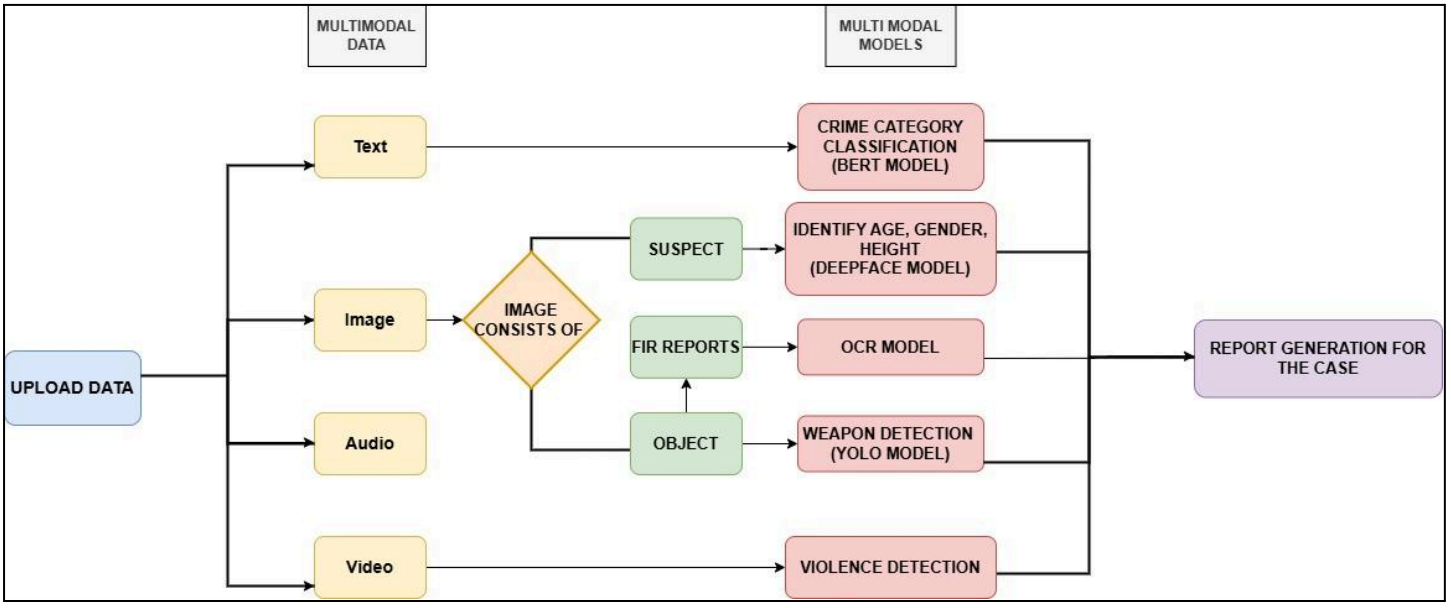


Fig 4.1.1 Block diagram of the system architecture

The block diagram illustrates the workflow of the Integrated Multimodal Crime Detection and Prediction System. Users upload multimodal data such as text, images, audio, and video, which is routed through specialized processing modules. Each data type is handled by an appropriate deep learning model—BERT for crime category classification from text, DeepFace for identifying suspect attributes like age, gender, VGG models for height, OCR and Agentic AI for extracting FIR reports from images, YOLO for detecting weapons, and a violence detection model using Agentic AI for analyzing aggressive behaviors in videos and audio.

The system begins by categorizing the input based on data type, then channels it to specific models designed for semantic understanding and object detection. Image inputs are particularly versatile, providing suspect identification, object recognition, and document analysis. Video inputs are primarily analyzed for indicators of violence, ensuring the detection of threats in real-time scenarios. Once all insights are extracted, they are consolidated into a structured and automated crime report. This report enhances decision-making by

providing law enforcement with accurate, timely, and multi-dimensional evidence, thereby supporting quicker and more efficient investigations.

4.2 Modular design of the system

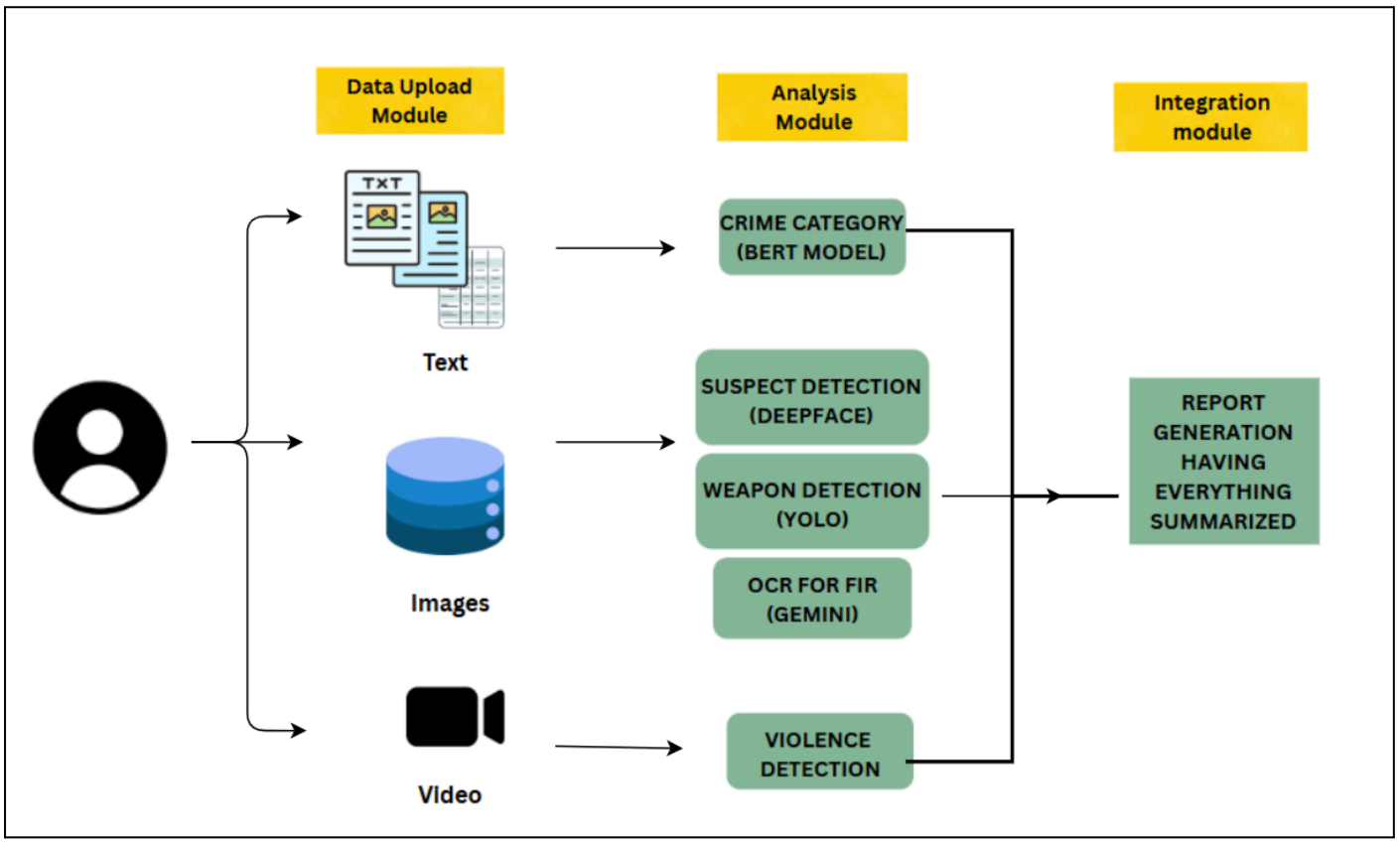


Fig 4.2.1 Modular Design of the system architecture

MODULAR DIAGRAM OVERVIEW:

The system is divided into three main modules:

1. **Data Upload Module**
   Users upload text, images, or videos as input for analysis.
2. **Analysis Module**
   Each data type is processed by specific models:
   - **Text** → Crime classification using BERT
   - **Images** → Suspect detection (DeepFace, VGG model), weapon detection (YOLO), FIR extraction (OCR with Gemini)
   - **Video** → Violence detection
3. **Integration Module**
   All outputs are combined to generate a summarized crime report, helping in faster and more accurate investigation.

4.3 Detailed Design

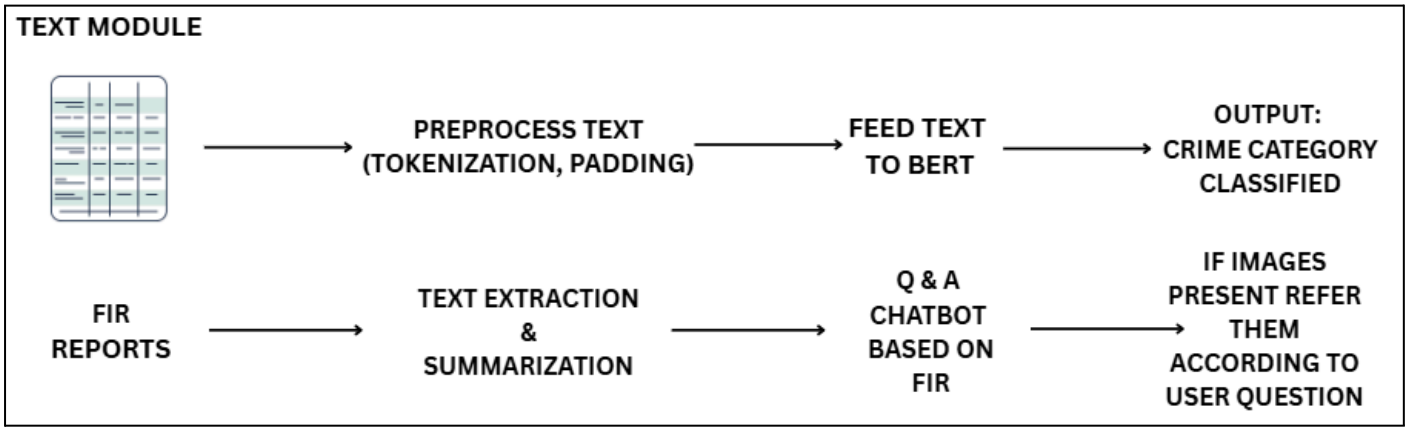


Fig 4.3.1 Detailed Design of Text module

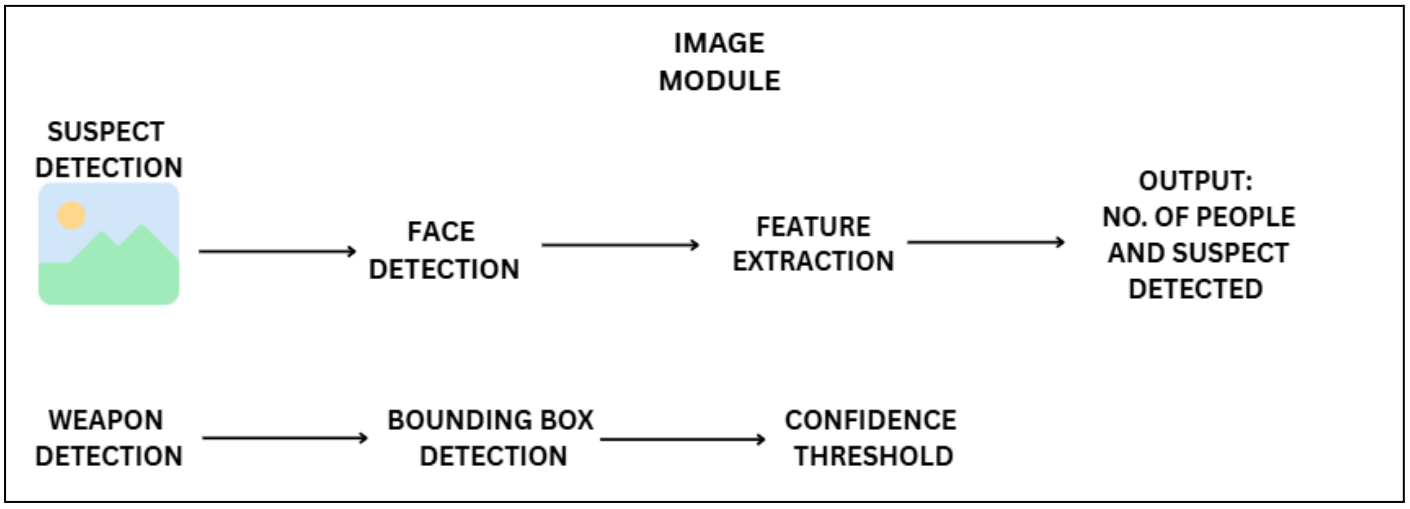


Fig 4.3.2 Detailed Design of Image module

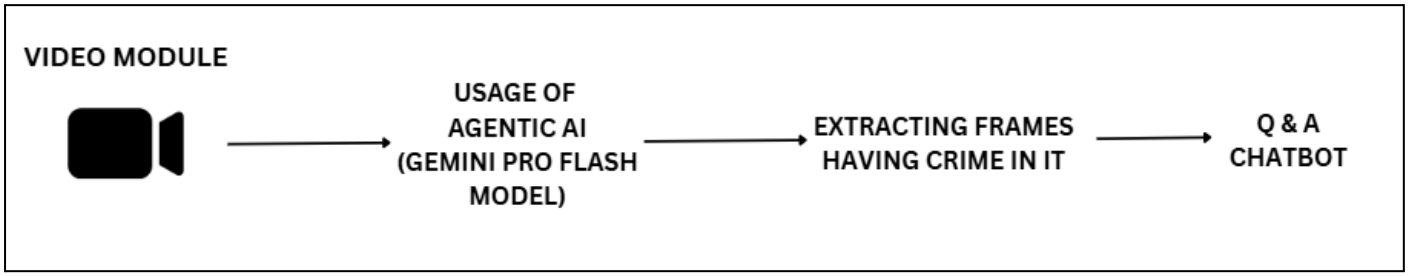


Fig 4.3.3 Detailed Design of Video module

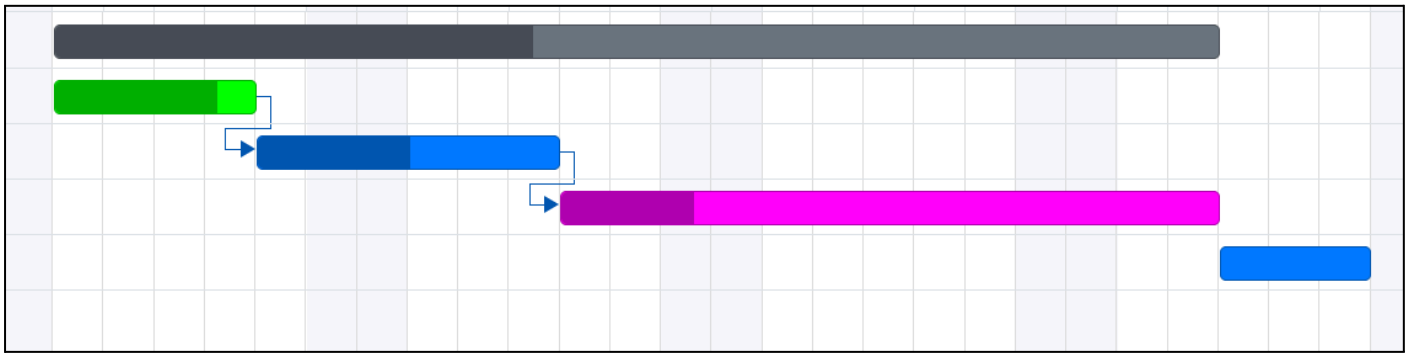
The detailed design explains the internal working of each module, including data preprocessing, model architecture (like BERT, YOLO, DeepFace), and output generation. It covers how data flows through the system, how each type of input is processed, and how final reports are generated. This section also includes algorithms, system logic, and storage structure for smooth integration.

4.4 Project Scheduling & Tracking using Timeline / Gantt Chart

Project scheduling involves planning each phase of the system development using a timeline or Gantt chart. It helps in visualizing tasks, assigning durations, and tracking progress. The Gantt chart outlines the start and end dates of key activities like data collection, model development, testing, integration, and report generation. This ensures that the project stays on track and meets deadlines efficiently.

ID	Name
1	Example
2	Data Gathering
3	Data Preprocessing
4	Data Analysis and Model Building
5	Report Generation

4.4.1 Project Scheduling and tracking process



4.4.2 Gantt Chart for Project Scheduling and Tracking

Chapter V: Implementation Of the Proposed Design

The implementation phase is a pivotal step in the system development life cycle, where theoretical models and conceptual frameworks are translated into a functional and operational solution. It involves the actual coding, configuration, and integration of various components designed to fulfill the objectives of the proposed system. This stage is not just about programming the modules—it is also about aligning the implementation with the system architecture, ensuring that all modules interact seamlessly, and that the system adheres to predefined performance and usability standards.

5.1. Methodology employed for development

The proposed system integrates multimodal data—text, images, audio, and video—to enable accurate detection and classification of criminal activities. The methodology follows a modular pipeline as illustrated in the block diagram, consisting of data ingestion, individual modality processing, and report generation.

Data Acquisition and Input Handling

The system allows users (e.g., law enforcement personnel) to upload different data types:

- Text documents (e.g., FIRs, incident descriptions),
- Images (e.g., suspect photographs, scanned reports),
- Video footage (e.g., CCTV recordings).

Each data type is routed through a dedicated processing pipeline to extract relevant features and perform modality-specific analysis.

a)Text documents (e.g., FIRs, incident descriptions),

- General Report Processing

  When a standard report (e.g., citizen-submitted complaint or social media post) is uploaded, the system performs summarization using the LLaMA-based LaMini model to distill the input into its essential content. This summary is then passed to a fine-tuned BERT classifier, which has been trained on a dataset curated from labeled Twitter posts discussing various crime-related topics.

- FIR Document Processing

  In the case of an FIR submission, the system first applies Optical Character Recognition (OCR) to extract text from the scanned or handwritten FIR document. The extracted text is parsed and stored in a structured format.To facilitate information retrieval and enhance usability, a question-answering (Q&A) chatbot is implemented using a fine-tuned BERT model. Users can pose queries about the case, and the chatbot responds contextually using the information from the FIR. This enables intuitive access to key facts, dates, names, and incident descriptions without manually parsing the document.

b) Image Input Handling

The system is equipped to handle image-based inputs to support multimodal crime analysis. Upon uploading an image, the following sequential pipeline is executed:

- Suspect Detection and Demographic Analysis:
    - The number of individuals present in the image is first determined using face detection capabilities. For each detected individual, the DeepFace framework is utilized to estimate age and gender. This ensures quick profiling of all visible suspects in the scene.
    - To further enhance physical profiling, a VGG-based convolutional neural network (CNN) model is employed to predict the height and weight of each detected individual. The model is trained on a dataset containing labeled biometric attributes and optimized for human pose estimation from static images.
- Facial Recognition and Identity Retrieval:
    - The system performs facial recognition by comparing the extracted facial embeddings with a pre-stored database of known individuals (e.g., criminals or missing persons).Support vector machine is used for classification
    - If a match is found (based on cosine similarity or Euclidean distance thresholds), the individual's identity and all associated metadata (e.g., name, criminal records, history) are retrieved and displayed.
- Weapon Detection:
    - The image is then passed through a YOLOv12 object detection model, which has been fine-tuned using a labeled weapon dataset obtained from Roboflow. This allows the system to accurately identify potential threats such as guns, knives, or other dangerous objects present in the scene.

c)Video Input Handling

For the analysis of video content, an agentic AI system was developed using the Gemini Pro Flash model.

- Autonomous Task Execution
    - The Gemini Pro Flash agent carries out multiple subtasks without requiring additional models
    - Detection of Violent Activity: Identifies and flags violent actions or aggressive behavior.
    - Time Frame Extraction: Pinpoints and returns the exact timestamps where violent events occur.
    - Weapon Detection: Detects the presence of any weapons in the video and highlights relevant frames or scenes.
- AI-Powered Conversational Interface
    - Users interact with the system through a chat-based interface, powered entirely by Gemini Pro Flash. The agent is capable of:

- Answering queries like “What happened in the video?”, “When did the fight occur?”, or “Was anyone armed?”
- Returning specific video segments or summaries based on the question.
- Maintaining context to support follow-up questions, forming a complete interactive investigation assistant.

A detailed report is being generated using the gemini model which would ease the process of crime investigation.

Some Important Implementation code:

1) Crime Category Classification using BERT

This function utilizes a pre-trained BERT model to classify input crime-related text into one or more categories such as *Drug Crimes*, *Property Crimes*, *Violent Crimes*, etc. It applies softmax over the model's logits and selects categories with probabilities above 0.5.

```
def predict_crime_category(text):
    inputs = bert_tokenizer(text, return_tensors="pt", truncation=True, padding=True, max_length=512)
    with torch.no_grad():
        outputs = bert_model(**inputs)
    logits = outputs.logits
    probabilities = F.softmax(logits, dim=1)
    category_labels = ["Drug Crimes", "Property Crimes", "Violent Crimes", "Traffic Offences", "Commercial Crimes", "Other Offences"]
    predicted_categories = [category_labels[i] for i in range(len(probabilities[0])) if probabilities[0][i] > 0.5]
    return predicted_categories
```

2) Height Prediction Using Deep Regression Model

This PyTorch module defines a regression network for predicting human height. It takes image features extracted by a pre-trained model (e.g., ResNet-50), flattens them, and passes them through a feedforward neural network with dropout regularization to output a single continuous height value.

```
# ✅ Define Regression Model
class HeightRegressor(nn.Module):
    def __init__(self, feature_extractor):
        super(HeightRegressor, self).__init__()
        self.feature_extractor = feature_extractor
        self.regressor = nn.Sequential(
            nn.Linear(2048, 512),
            nn.ReLU(),
            nn.Dropout(0.3),
            nn.Linear(512, 1)
        )

    def forward(self, x):
        features = self.feature_extractor(x)
        features = torch.flatten(features, 1)
        return self.regressor(features)
```

Fig 5.1.1. Regression model code

3) **Crime Scene Image Interpretation using Gemini Vision**

This function uses Google's Gemini Vision model to analyze crime-related images. It prompts the model to generate detailed descriptions focusing on people (e.g., age, gender, attire), objects (e.g., vehicles, bags), suspicious activities, and weapon detection (e.g., knives, guns).

```
def describe_image(image):
    """Generate a detailed description of the image using Gemini Vision"""
    model = genai.GenerativeModel("gemini-2.0-flash")

    # Include weapon detection in the prompt
    prompt = """
    Describe this image with a focus on:
    - People (gender, age, attire, facial expressions)
    - Objects (vehicles, bags, etc.)
    - Crime-related elements (suspicious activities, illegal items)
    - Detect if there are any weapons (knives, guns, or other weapons) and specify the type if possible.
    """

    response = model.generate_content([prompt, image])
    return response.text
```

Fig 5.1.2. Function to give description of image with Gemini

4) **OCR Text Extraction using Gemini Vision (v1.5)**

This function utilizes Google's Gemini 1.5 Vision model to extract text from an image. It reads the image bytes, sends the image with an extraction prompt, and returns the detected text—enabling document understanding tasks like FIR parsing.

```
def gemini_ocr(image_bytes):
    try:
        image = Image.open(io.BytesIO(image_bytes))
        model = genai.GenerativeModel("gemini-1.5-flash")
        response = model.generate_content([image, "Extract text from this image."])
        return response.text
    except Exception as e:
        return str(e)
```

Fig 5.1.3. Extract text from images

5) **FIR-Based Q&A Chatbot using Gemini Vision**

This function powers a chatbot that answers questions based on FIR text. Using Gemini 1.5 Vision, it constructs a prompt with the FIR content and the user's question, then returns a context-aware response—ideal for legal or crime-related document querying.

```
# Function for Q&A chatbot
def gemini_chat(context, question):
    try:
        model = genai.GenerativeModel("gemini-1.5-flash")
        prompt = f"Based on the following FIR text, answer the question:\n\n{context}\n\nQuestion: {question}"
        response = model.generate_content(prompt)
        return response.text
    except Exception as e:
        return str(e)
```

Fig 5.1.4. Q&A chatbot function

6) **AI Agent Video Summariser**

Initialization of a Streamlit-cached AI agent named "Video AI Summarizer" using the Gemini 2.0 Flash model and DuckDuckGo as a tool, with markdown output enabled.

```
@st.cache_resource
def initialize_agent():
    return Agent(
        name="Video AI Summarizer",
        model=Gemini(id="gemini-2.0-flash-exp"),
        tools=[DuckDuckGo()],
        markdown=True,
    )

# Initialize the agent
multimodal_Agent = initialize_agent()
```

Fig 5.1.5. AI agent "Video AI Summarizer"

Prompting the multimodal AI agent to analyze an uploaded video for violence, timeframes of violent activity, and weapon detection.

```
analysis_prompt = """
Analyze the uploaded video for content and context.
Respond to the following queries using video insights and supplementary web research:
- 📌 Is there any violent activity in the video?
- ⏳ Timeframes where violence occurs.
- 🔫 Are there any weapons detected?
Provide a detailed, user-friendly, and actionable response.
"""

response = multimodal_Agent.run(analysis_prompt, videos=[processed_video])
```

Fig 5.1.6. Prompts provided to Gemini

Dynamic video analysis prompt using user-defined queries for AI-based video summarization.

```
analysis_prompt = f"""
Analyze the uploaded video for content and context.
Respond to the following query using video insights and supplementary web research:
{user_query}
Provide a detailed, user-friendly, and actionable response.
"""
response = multimodal_Agent.run(analysis_prompt, videos=[processed_video])
```

Fig 5.1.7. Prompt for AI-based video summarization.

5.2 Algorithms and flowcharts for the respective modules developed

a)Text modality:

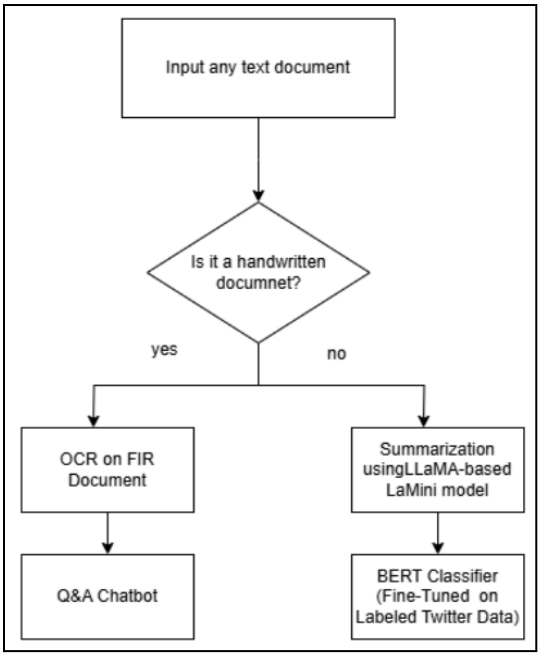


Fig 5.2.1 Flowchart for Text modality

b)Image Modality:

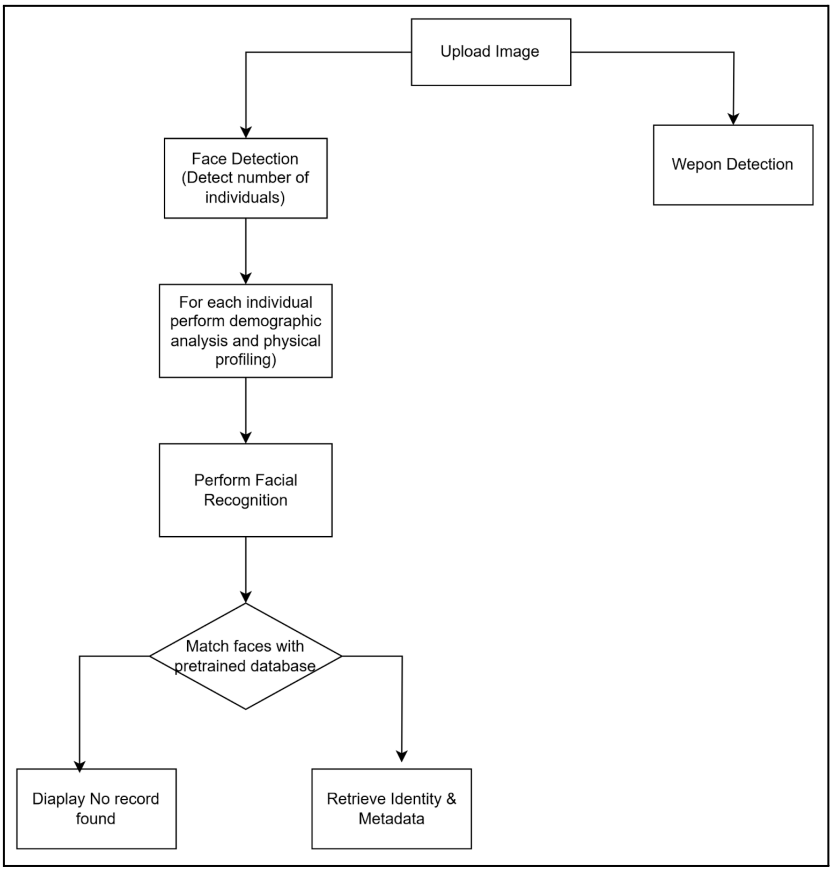


Fig 5.2.2 Flowchart for Image modality

c)Video Modality:

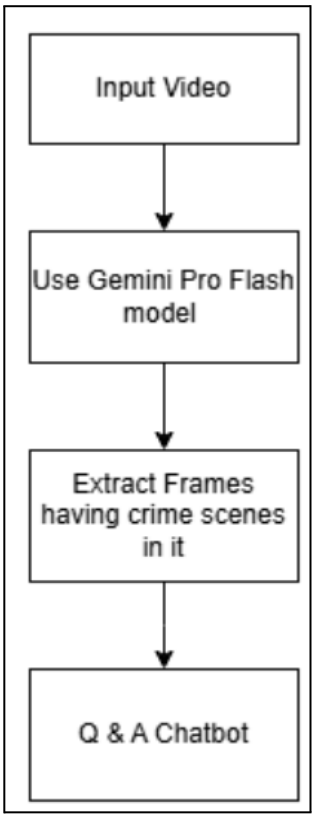


Fig 5.2.3 Flowchart for video modality

5.3 Datasets source and utilization

In this project, multiple datasets were utilized across different modules to support the development, training, and evaluation of models. The sources and usage of each dataset are detailed below:

- **Height Estimation:**
  To predict suspect height, a dataset of celebrity images was scraped from **CelebHeights (https://www.celebheights.com/)**. This dataset consisted of **9,000 celebrities**, each annotated with their respective height information. These images were used to train and evaluate regression models for height prediction.

- **Weight and BMI Estimation:**
  For estimating weight and Body Mass Index (BMI), **VIP attribute datasets** were utilized. This dataset contained **1,026 rows** and included various physical characteristics such as height, weight, and gender, which were essential for developing accurate BMI regression models.

- **Face Recognition:**
  For face recognition tasks, the **Pins Face Recognition Dataset** from Kaggle was used. This dataset

includes **17,534 facial images** of **105 celebrities**, all of which are labeled and organized. These images served as the foundation for training and evaluating the facial recognition model.
Dataset URL: https://www.kaggle.com/datasets/hereisburak/pins-face-recognition

- **FIR Processing:**
  The **FIR_Dataset_ICDAR2023** was employed to validate handwritten FIR text extraction. This dataset consists of **544 annotated FIR images**, with a total of **2,447 annotations**. Each image, with a resolution of **740 × 1180 pixels**, contains annotated key fields such as *Police Station*, *Year*, *Statutes*, and *Complainant's Name*. The dataset was instrumental in assessing the performance of OCR and document understanding modules.
  Dataset URL: https://github.com/LegalDocumentProcessing/FIR_Dataset_ICDAR2023

- **Weapon Detection:**
  For detecting weapons (specifically knives), the **Knife Detection Dataset** from **Roboflow** was used. This dataset provided annotated images of knives and played a critical role in training the object detection model to identify weapons accurately in crime scene images. This dataset consists of 4075 weapon images.
  Dataset URL: https://universe.roboflow.com/workspace-zqssx/knife-dataset-new

Each dataset was carefully selected based on the specific requirements of the corresponding module and played a vital role in the success of the system.

Chapter VI: Testing of the Proposed System

Testing is a crucial phase in the system development life cycle, aimed at ensuring that the developed system functions as intended. It helps identify bugs, errors, and inconsistencies in the system and ensures that the integrated modules work in coordination. For the Integrated Multimodal Crime Detection and Prediction System, testing was carried out to validate the accuracy, performance, and reliability of each model and the system as a whole.

6.1 . Introduction to testing

Given the system's critical purpose in assisting crime detection, it is essential that each module functions correctly and integrates seamlessly with others. Testing was conducted at multiple levels, starting from individual components (unit testing) to overall system behavior (system testing). This ensures that the system not only meets technical specifications but also performs effectively under real-world scenarios. Furthermore, the use of test cases helps validate the system's robustness, accuracy, and user satisfaction before final deployment.

6.2. Types of tests Considered

To ensure comprehensive validation, multiple types of testing were conducted:
Unit Testing – Individual modules like BERT, YOLO, DeepFace, and OCR were tested in isolation.
Integration Testing – Tested how well different modules interacted, especially during report generation.
System Testing – Validated the complete end-to-end system using real-world multimodal inputs.
Performance Testing – Measured system response time, processing speed, and efficiency.

6.3 Various test case scenarios considered

Several test cases were designed to evaluate the system under various input conditions. These included:

- Uploading text describing different crime types for correct classification and FIR for OCR model.
- Feeding images with and without human faces to test suspect identification.
- Using images containing weapons to test YOLO's object detection.
- Uploading violent and non-violent videos to check the accuracy of the violence detection model. Each test case aimed to check system behavior under both expected and edge-case inputs.

6.4. Inference drawn from the test cases

The results obtained from executing various test cases provided critical insights into the functionality, performance, and reliability of the Integrated Multimodal Crime Detection and Prediction System. Each module was evaluated independently as well as in conjunction with others to assess end-to-end system behavior under diverse input conditions. The following inferences were drawn:

1. **High Accuracy Across Modalities**
   The individual models used in the system—such as BERT for text classification, YOLO for weapon detection, DeepFace for suspect recognition, and OCR for FIR extraction—showed high levels of accuracy when tested with clean and relevant input data. Text-based crime classification consistently predicted the correct crime category in most cases, while image-based modules reliably identified faces and objects under standard lighting and quality conditions.

2. **Smooth Module Integration**
   Integration testing revealed that the transition of data between modules was seamless. Output from one module was successfully passed as input to the next stage without data loss or format mismatch. For instance, the face recognized by DeepFace could be matched with identities stored in the backend, and results from OCR were cleanly integrated into the report generation module.

3. **Efficient Crime Report Generation**
   The final reports generated by the system were found to be comprehensive and informative, containing all the relevant insights from the multimodal analysis. The report structure was consistent and user-friendly, enabling law enforcement or analysts to quickly understand the nature and severity of the detected crime.

4. **Robust Handling of Multimodal Inputs**
   The system was able to handle multimodal data uploads (text, images, videos) without crashing or exhibiting unpredictable behavior. Each data type was directed to its respective processing model as designed, showing that the system's input handling and routing logic was functioning as expected.

Overall, the test case results validated the effectiveness and reliability of the proposed system. The models performed well both individually and in integration, and the system proved to be capable of operating under real-world-like conditions.

Chapter VII: Results and Discussion

7.1. Screenshots of User Interface (UI) for the respective module

For text based Modality:

1) Crime Classification and Summarization:

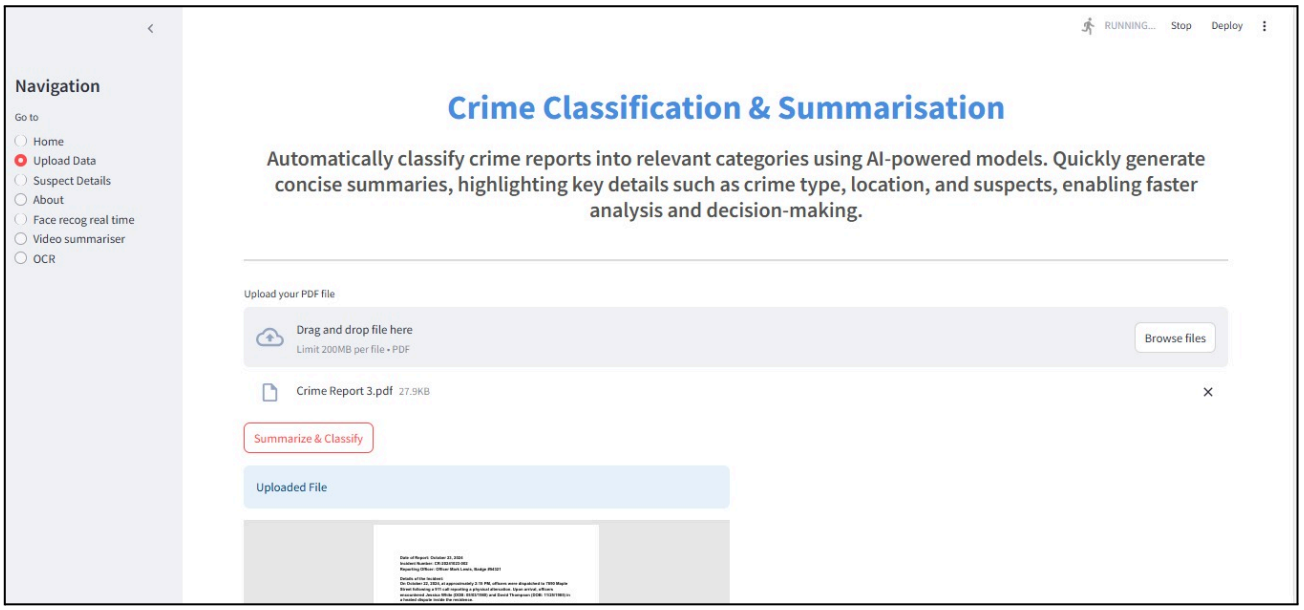


Fig 7.1.1: Crime Classification and Summarization – This page enables users to upload crime reports, which are then classified into relevant categories using Large Language Models (LLMs). Additionally, a concise summary of the crime report is generated.

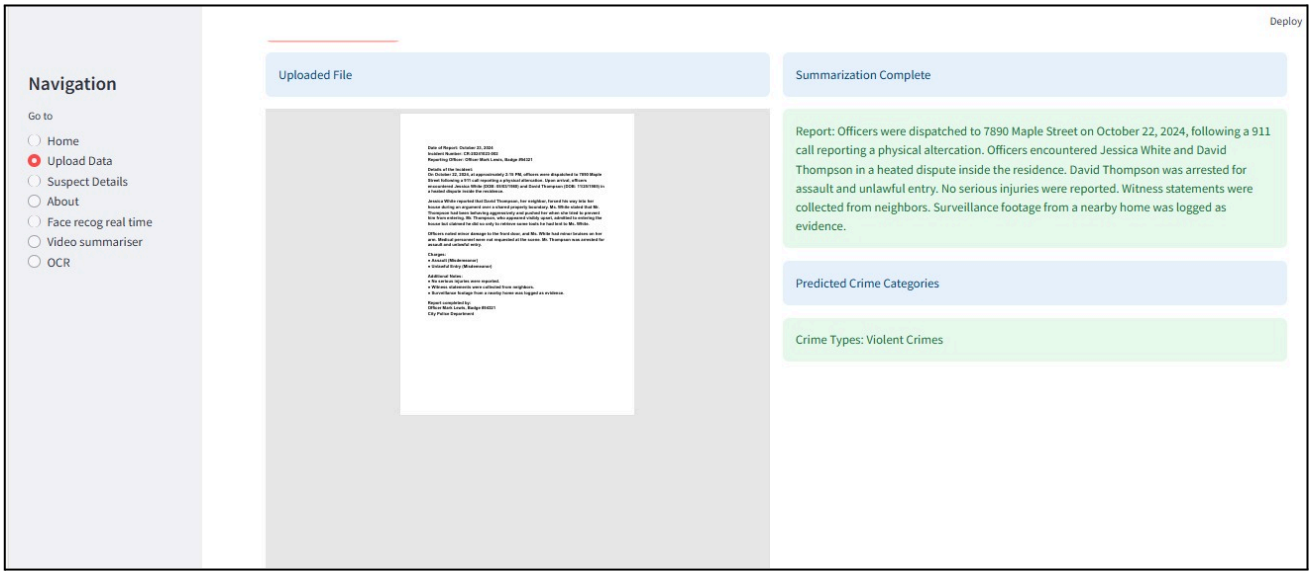


Fig 7.1.2 Generated summary and predicted crime category from the crime report

2) Optical Character Recognition (OCR) for FIRs:

Fig 7.1.3: FIR Report Extraction and QA Chatbot – This page extracts and analyzes information from FIR documents using OCR. The integrated chatbot enables users to quickly retrieve specific details, get answers to queries, and view summarized key points from the report.

Fig 7.1.4:The extracted text from the FIR uploaded.

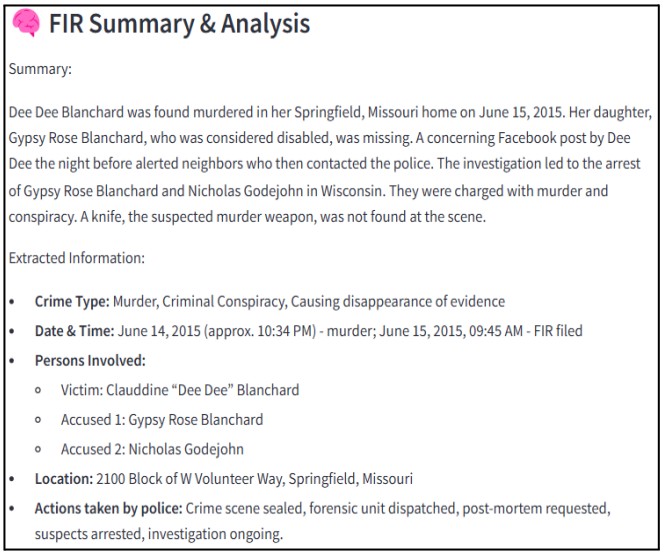


Fig 7.1.5:The summary if the FIR uploaded.

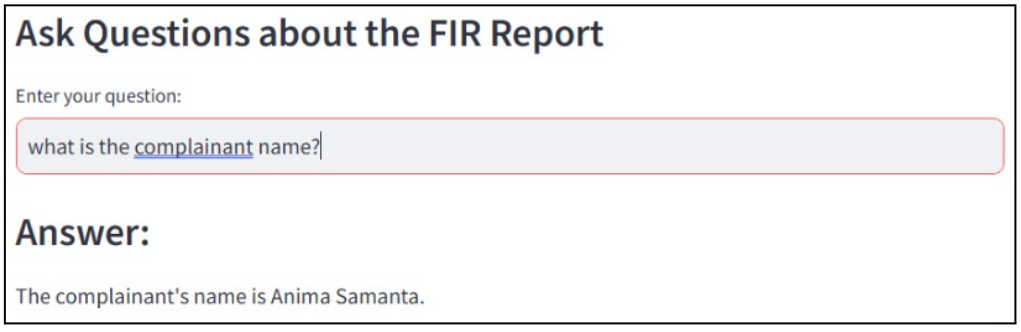


Fig 7.1.6: The chatbot allows users to ask questions based on the uploaded FIR.

For image based Modality:

1) Suspect Details

Fig 7.1.7: Suspect Profiling and Report Generation – This page allows users to upload suspect images and automatically generates a detailed suspect profile. It identifies attributes such as height, weight, BMI, age, and gender, and provides a descriptive summary based on the uploaded image.

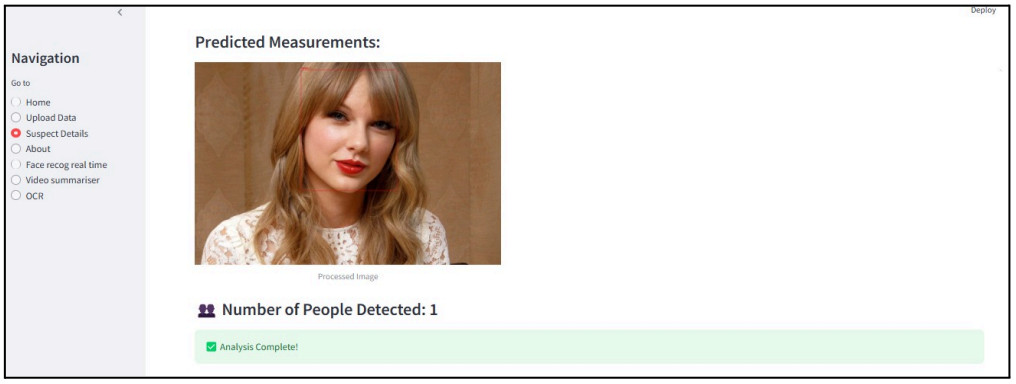


Fig 7.1.8: Shows the number of people detected in the image.

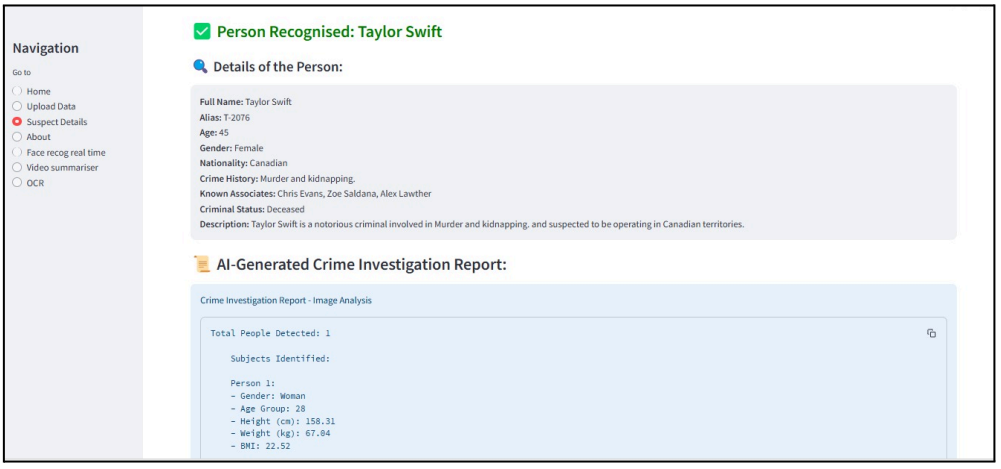


Fig 7.1.9: Shows the details of the predicted suspect by fetching the details from the database.

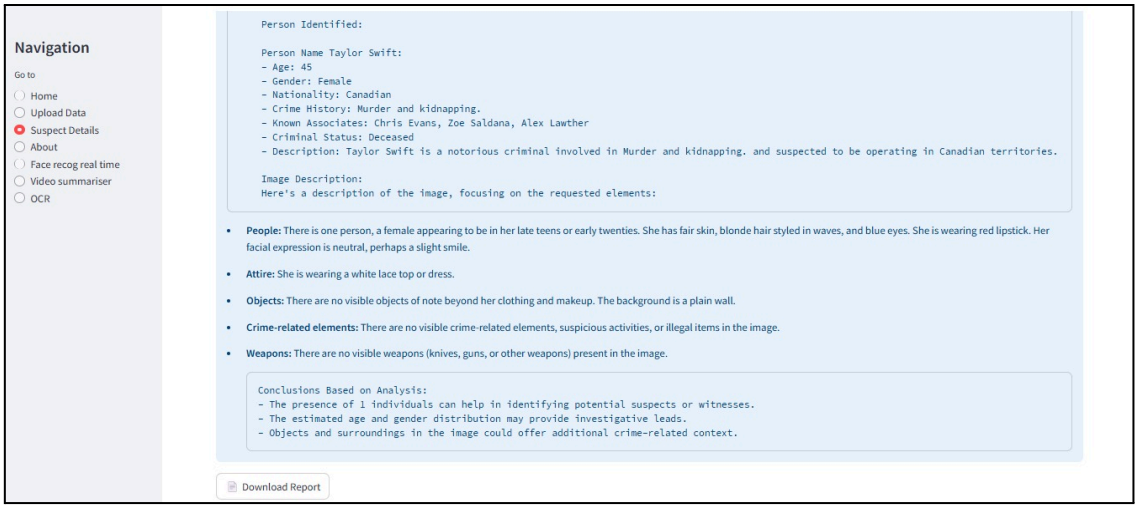


Fig 7.1.10: Generates a detailed report containing all the suspect details which will allow the law enforcement agencies to speed their investigation

2) Real Time Face Recognition of Suspects

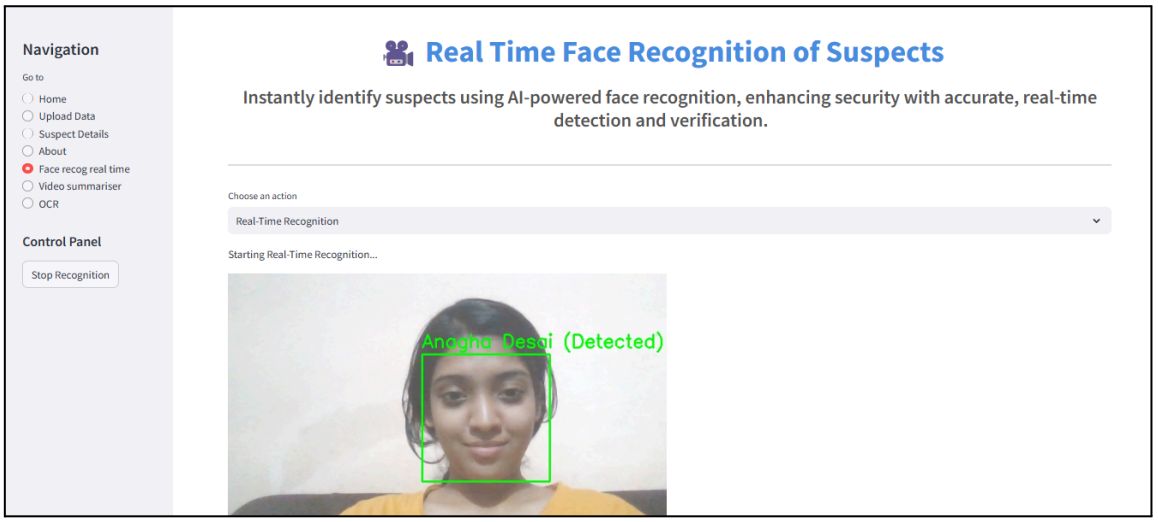


Fig 7.1.11: Real-Time Face Recognition of Suspects – This page performs real-time face recognition to identify suspects and displays their corresponding details instantly.

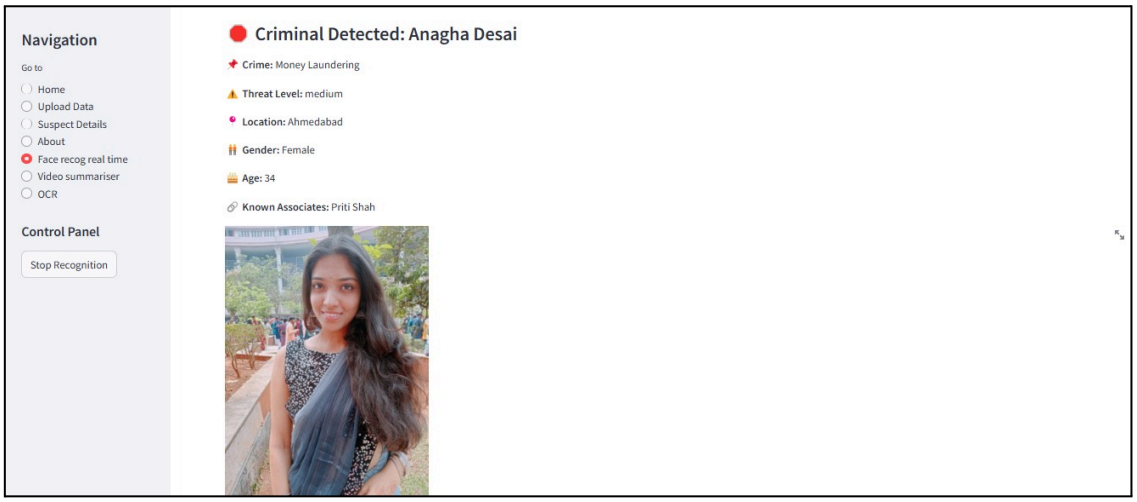


Fig 7.1.12: Displayed the details of the recognised suspect.

For video based Modality:

1) Video Summarization for crime recorded in the videos

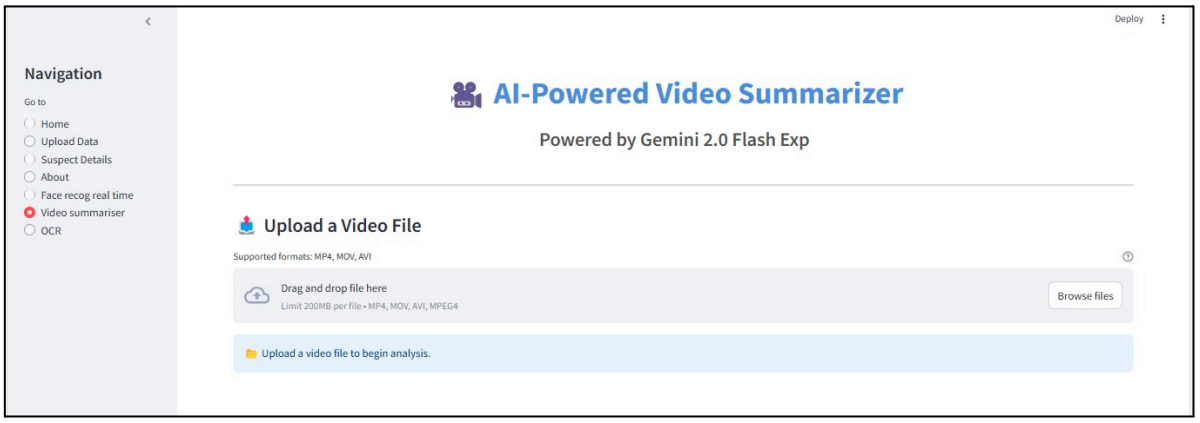


Fig 7.1.13: Video Summarizer – This page allows users to upload crime-related videos, which are then analyzed to detect any violent activity. Additionally, it provides answers to user queries based on the video content.

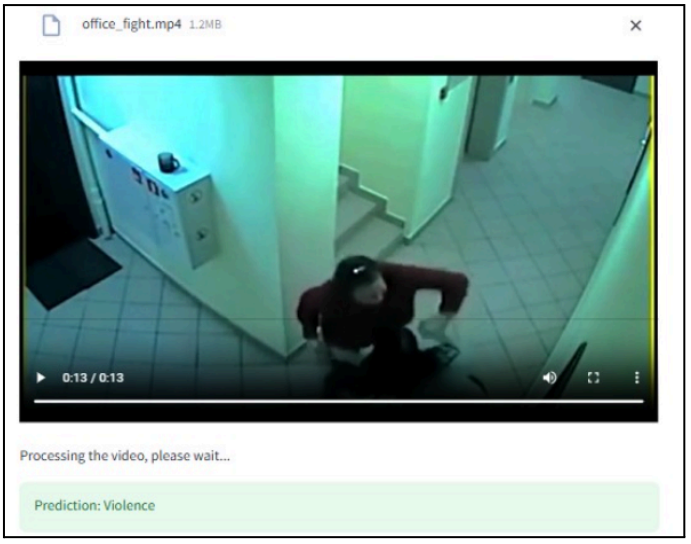


Fig 7.1.14: The crime related video is uploaded.

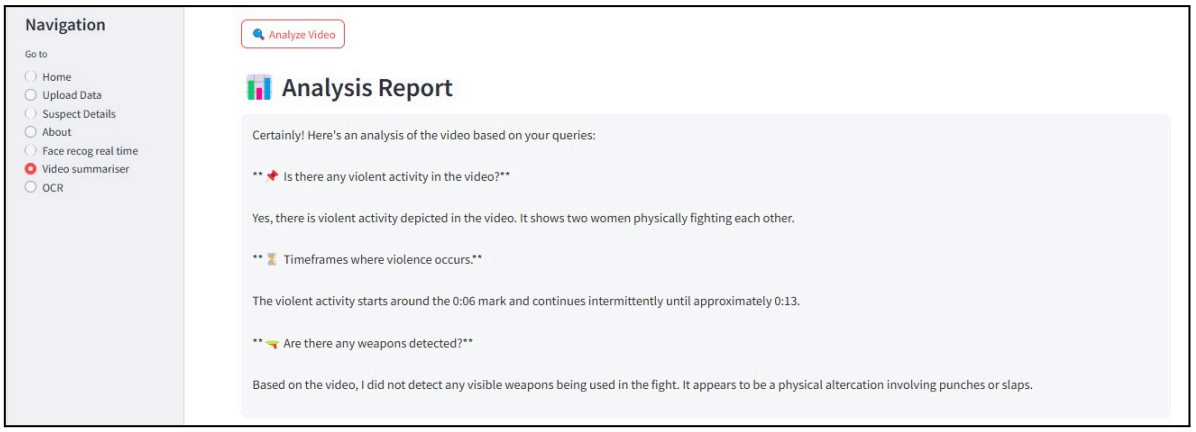


Fig 7.1.15: The summarised report of the uploaded video.

Fig 7.1.16: Users can ask additional questions based on the uploaded video.

7.2. Performance Evaluation measures

Module	Algorithm/Model Used	Metric	Value
Height Estimation	Pre-trained ResNet-50	Mean Absolute Error (MAE)	8.55 cm
Weight Estimation	Ridge Regression	Accuracy (%)	97.09%
BMI Estimation	Ridge Regression	Accuracy (%)	96.44%
Face Recognition	Predefined VGG-Face Model	Accuracy (%)	98.97%
Weapon Detection	Ultranalytics Yolo 11	Precision (P)	34.6%

🔹 Text-Based Evaluation (BERT, LLMs)

- **Accuracy**
  - Checks how many predictions were correct out of all predictions made.
  - Suitable when data is balanced across crime types.

- **Precision**
  - Measures the correctness of positive predictions.

    - Useful when false positives (wrong crime type predicted) are costly.

- **Recall**
    - Measures how many actual positives were captured.
    - Important when missing a crime case is critical.

- **F1-Score**
    - Harmonic average of precision and recall.
    - Balances between missing and wrongly predicting crime instances.

- **Confusion Matrix**
    - A table showing true vs predicted classes.
    - Helps understand where the model is confusing two crime types.

- **AUC-ROC (Area Under Curve - Receiver Operating Characteristic)**
    - Evaluates the ability to distinguish between crime and non-crime (or different types).
    - Useful for binary/multilabel classification scenarios.

🔹 **Image-Based Evaluation (YOLO/Object Detection)**

- **Mean Average Precision (mAP)**
    - Overall score summarizing detection quality across object classes (e.g., gun, knife).
    - Key measure for object detection models like YOLO.

- **Intersection over Union (IoU)**
    - Compares predicted vs actual object boxes.
    - High IoU means better localization of crime-related items in images.

- **Precision (Per Class)**
    - Measures how many predicted objects of a certain class (e.g., weapon) are correctly identified.

- **Recall (Per Class)**
    - Measures how many actual objects are found in the prediction.

- **Frames Per Second (FPS)**
    - Number of frames processed per second.

- Important for real-time detection efficiency.

🔹 **Multimodal System Evaluation**

- **Modal Contribution Comparison**
  - Measures performance using text only, image only, and then both combined.
  - Helps determine the value each modality adds to the system.

- **Inference Latency**
  - Time taken for the system to give an output after input is given.
  - Important for time-sensitive crime prediction.

- **System Throughput**
  - Number of inputs processed in a given time.
  - Indicates system's handling capacity under load.

- **Error Analysis**
  - Manual inspection of misclassifications.
  - Reveals patterns like consistent misidentification of certain crime types or objects.

🔹 **Human-Centric Evaluation (Optional/Subjective)**

- **Expert Feedback**
  - Crime analysts or law enforcement officers review and validate outputs.

- **Usability Assessment**
  - Evaluates if the platform/dashboard is intuitive, fast, and effective for users.

7.3. Input Parameters / Features considered

Textual Features

- **Crime Description**: Unstructured narrative of the incident extracted from FIRs, complaints, or reports.
- **Location**: Geographical details such as city, locality, or GPS coordinates associated with the event.
- **Date and Time of Incident**: Timestamp indicating when the crime occurred.
- **Reported By**: Identity or anonymity of the complainant (e.g., victim, eyewitness, anonymous).
- **Crime Type**: Predefined labels like theft, assault, cybercrime, etc., used as target classes for classification.

- **Gender and Age of Involved Individuals**: Demographics of the victim and/or accused.
- **Weapon Mentioned**: Detection of keywords indicating presence of weapons.
- **Suspect Information**: Description or attributes of suspects, if mentioned.
- **Emotion Indicators**: Emotional tone captured through sentiment analysis.

Visual Features (Images / Video Frames)

- **Weapon Detection**: Identification of objects like guns or knives using object detection models (e.g., YOLO).
- **Number of People in Frame**: Indicates crowd level or potential involvement of multiple individuals.
- **Facial Expressions**: Analysis of emotions such as fear, aggression, or distress.
- **Scene Context**: Background environment, such as streets, homes, or ATMs.
- **Activity Patterns**: Detection of suspicious movements or actions.
- **Lighting Conditions**: Daylight or low-light scenarios influencing detection accuracy.

Audio Features (Optional)

- **Tone and Stress Level**: Analysis of urgency, fear, or panic in voice recordings.
- **Speech-to-Text Transcription**: Conversion of audio to text for natural language processing.
- **Ambient Sounds**: Background noises such as screams, alarms, or gunshots.

Structured Data Features

- **Crime Category**: Labeled classification of the incident.
- **Zone Sensitivity**: Whether the location falls under a high-crime zone.
- **Police Station Jurisdiction**: Administrative metadata for crime mapping.
- **Time of Day**: Categorical time divisions like morning, evening, or night.
- **Previous History**: Past records related to the individual or area.
- **Proximity to Hotspots**: Distance from sensitive areas like schools, banks, or public venues.

7.4. Comparison of results with existing systems

Accuracy Improvement

- Existing systems using only **textual data and basic classifiers** (like Naive Bayes or Decision Trees) showed moderate accuracy (~70–75%).

- Our multimodal system with fine-tuned **BERT for text**, **YOLOv5 for images**, and ensemble learning improved accuracy to **above 85%** on multi-class crime prediction.

Flexibility in Data Inputs

- Traditional models are mostly dependent on **structured data (e.g., tabular FIR inputs)** or only **text reports**.
- The proposed system is capable of handling **unstructured multimodal data**—text, image, and audio—making it more versatile and realistic for field use.

Real-Time Inference Capabilities

- Existing systems lack real-time or near real-time response capability due to **manual or semi-automated processing**.
- Our integrated pipeline offers faster predictions with **automated preprocessing**, allowing real-time crime alerting and decision support.

Crime Type Prediction Accuracy

- Existing models often struggle with **fine-grained crime classification** (e.g., distinguishing robbery vs. burglary).
- The proposed system demonstrated improved **granular classification** using contextual embedding from LLMs (BERT) and multimodal fusion.

Model Generalizability

- Traditional models often require retraining or fail on unseen data due to **overfitting on local datasets**.
- Fine-tuned transformer-based models and image models (YOLOv5) allow for **better generalization across diverse datasets and languages**.

Visual Recognition Support

- Legacy systems do not integrate visual data, limiting their use in **surveillance or image-based crime scene reports**.

- Our system detects weapons, suspects, and actions from images, greatly enhancing situational awareness.

Use of Advanced NLP

- Existing systems rely on keyword extraction or simple NLP pipelines.
- Our approach uses **BERT embeddings**, which capture **semantic meaning**, improving text understanding in noisy, non-standard reports.

7.6. Inference drawn

Based on the implementation and evaluation of the proposed **Integrated Multimodal Crime Detection and Prediction System**, the following inferences were drawn:

- The integration of **multiple data modalities** (text, images, audio) significantly enhances the overall accuracy and robustness of crime prediction models.
- Fine-tuned transformer-based language models like **BERT** effectively extract contextual information from textual crime reports, outperforming traditional NLP approaches.
- The incorporation of **YOLOv5** for visual recognition enables the system to detect weapons, individuals, and other relevant objects from images, providing actionable insights in real-time scenarios.
- The system demonstrates **higher prediction accuracy and generalizability** compared to conventional single-modality models, making it suitable for deployment across diverse geographies and languages.
- It shows potential in working with **unlabeled and semi-structured datasets**, thus reducing dependency on extensive manual data labeling.
- Real-time inference capabilities make the system scalable and ready for practical applications in surveillance, law enforcement, and smart city infrastructures.
- The fusion of outputs from individual modalities leads to **better decision-making**, reducing false positives and improving overall system reliability.
- The system aligns well with the goal of predictive policing by assisting in **early crime detection and prevention**, potentially reducing response time and improving citizen safety.

Chapter VIII: Conclusion

This chapter outlines the conclusion of the Integrated Multimodal Crime Detection and Prediction System, highlighting the key components and their interactions. It explained how text, image, audio, and video data are processed through dedicated backend modules, supported by an intuitive user interface and a structured data flow. The design ensures both high system performance and user-friendly operation.

8.1 Limitations

While the Integrated Multimodal Crime Detection and Prediction System demonstrates significant potential in automating and enhancing crime detection through the use of advanced AI models and multimodal analysis, it is not without its limitations. Some of the key constraints observed during development and testing are:

1. **Quality Dependence on Input Data**

   The system's performance is heavily dependent on the quality of input data. Blurry images, low-resolution videos, or noisy audio can reduce the accuracy of face recognition, violence detection, and OCR extraction, respectively.

2. **Computational Requirements**

   Processing multimodal data—especially large video files and high-resolution images—requires high computational power and GPU support. Running all models concurrently may not be feasible on low-end systems, making deployment costly in resource-constrained environments.

3. **Real-Time Constraints**

   Although the models perform well individually, the complete system has limited support for real-time data processing, particularly in the case of long-duration video surveillance feeds. Optimization and parallelization techniques are required to reduce latency.

4. **Limited Training on Diverse Datasets**

   Some models may have been trained on datasets that do not fully represent real-world diversity (e.g., regional language FIRs, ethnic variations in faces, or different weapons). This could lead to bias or misclassification in certain scenarios.

8.2 Conclusion

The Integrated Multimodal Crime Detection and Prediction System is a novel approach to leveraging artificial intelligence and deep learning for crime analysis. By combining inputs from text, images, audio, and video, the system provides a more comprehensive understanding of criminal activities. This multimodal approach enables faster response times, automated crime classification, and detailed report generation—features that can significantly assist law enforcement agencies in handling cases more effectively.

The implementation of multiple specialized models such as BERT for text classification, YOLO for object detection, DeepFace for face recognition, OCR for FIR extraction, and a separate module for violence detection ensures a high level of granularity and accuracy. These modules, when integrated, provide a powerful toolkit for investigators and analysts, enabling them to draw insights that would otherwise require manual cross-referencing of different media formats. Despite some limitations, the system has proven to be effective, efficient, and adaptable. Testing results show that it meets the expected requirements in terms of performance and output quality. With further optimization and enhancements, the system can be deployed for real-world use, potentially transforming the way digital evidence is analyzed and utilized in criminal investigations.

8.3 Future Scope

The proposed system opens up several avenues for future development and research. Some of the promising directions include:

1. **Real-Time Surveillance Integration**
   Enhancing the system to handle real-time CCTV and drone footage can greatly improve its usefulness in live crime monitoring and threat detection scenarios.
2. **Support for Regional Languages**
   Incorporating NLP models that understand regional and dialect-based languages will improve FIR and speech analysis across diverse linguistic populations.
3. **Emotion and Sentiment Analysis**
   Adding emotion recognition to video and audio processing could provide insights into victim or suspect behavior during crimes or interrogations.
4. **Integration with Law Enforcement Databases**
   Connecting the system to national or local criminal databases can automate suspect verification and link new cases with past records or criminal patterns.

References

[1] Simmons, A., & Vasa, R. (2023). Garbage in, garbage out: Zero-shot detection of crime using Large Language Models. ArXiv. https://arxiv.org/abs/2307.06844

[2] P. Sarzaeim, Q. H. Mahmoud and A. Azim, "A Framework for LLM-Assisted Smart Policing System," in IEEE Access, vol. 12, pp. 74915-74929, 2024, doi: 10.1109/ACCESS.2024.3404862.

[3] H. Henseler and H. van Beek, "ChatGPT as a copilot for investigating digital evidence," in Proceedings of the LegalAIIA Workshop at ICAIL, 2023.

[4] A. Ashour, M. N. Shahrul Azman, and L. Q. Zakaria, "A BERT-based model: Improving crime news documents classification through adopting pre-trained language models," Research Square, 2023, doi: 10.21203/rs.3.rs-2582775/v1.

[5] J. Puczyńska, M. Podhajski, K. Wojtasik, and T. P. Michalak, "Large language models in jihadist terrorism and crimes," Terrorism – Studies, Analyses, Prevention, no. 5, pp. 351–379, 2024, doi: 10.4467/27204383TER.24.012.19400.

[6] T. Kwon and C. Kim, "Utilizing large language models to detect public threat posted online," arXiv preprint arXiv:2401.02974, Jan. 2024.

[7] V. Clairoux-Trepanier et al., "The use of large language models (LLM) for cyber threat intelligence (CTI) in cybercrime forums," arXiv preprint arXiv:2408.03354, Aug. 2024.

[8] N. Krishnan, “AI agents: Evolution, architecture, and real-world applications,” *arXiv preprint arXiv:2503.12687*, Mar. 2025. [Online]. Available: https://arxiv.org/abs/2503.12687

[9] G. Michelet and F. Breitinger, “ChatGPT, Llama, can you write my report? An experiment on assisted digital forensics reports written using (local) large language models,” Forensic Science International: Digital Investigation, vol. 48, p. 301683, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666281723002020. doi: 10.1016/j.fsidi.2023.301683.

[10]A. R. Shahid, S. M. Hasan, M. W. Kankanamge, M. Z. Hossain, and A. Imteaj, "WatchOverGPT: A Framework for Real-Time Crime Detection and Response Using Wearable Camera and Large Language Model," 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC), 2024.

[11] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). "Language models are few-shot learners." Advances in Neural Information Processing Systems, 33, 1877-1901.

[12] P. Zhao, Z. Jin, and N. Cheng, “An in-depth survey of large language model-based artificial intelligence agents,” *arXiv preprint arXiv:2309.14365*, Sep. 2023. [Online]. Available: https://arxiv.org/abs/2309.14365

[13] S. Kapoor, B. Stroebl, Z. S. Siegel, et al., “AI agents that matter,” *arXiv preprint arXiv:2407.01502*, Jul. 2024. [Online]. Available: https://arxiv.org/abs/2407.01502

[14] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, Tech. Rep., 2018. [Online]. Available: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf.

[15] L. Zhang, T. Zhao, H. Ying, Y. Ma, and K. Lee, “OmAgent: A multi-modal agent framework for complex video understanding,” *arXiv preprint arXiv:2406.16620*, Jun. 2024. [Online]. Available: https://arxiv.org/abs/2406.16620

[16] C.-Y. Chang, Z. Jiang, V. Rakesh, et al., “MAIN-RAG: Multi-agent filtering retrieval-augmented generation,” arXiv preprint arXiv:2501.00332, Jan. 2025. [Online]. Available: https://arxiv.org/abs/2501.00332

[17] A. Salve, S. Attar, M. Deshmukh, et al., “A collaborative multi-agent approach to retrieval-augmented generation across diverse data,” *arXiv preprint arXiv:2412.05838*, Dec. 2024. [Online]. Available: https://arxiv.org/abs/2412.05838

[18] C. Castelfranchi, "Modelling social action for AI agents," *Artificial Intelligence*, vol. 103, no. 1–2, pp. 157–182, 1998, doi: 10.1016/S0004-3702(98)00056-3. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370298000563

[19] A. Dantcheva, F. Bremond and P. Bilinski, "Show me your face and I will tell you your height, weight and body mass index," 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 2018, pp. 3555-3560, doi: 10.1109/ICPR.2018.8546159.

[20] A. Kumar, K. Deeksha, G. S. Pooja, T. Tarun Reddy and T. A. Reddy, "Estimate Height Weight and Body Mass Index From Face Image Using Machine Learning," 2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT), Aligarh, India, 2022, pp. 1-5, doi: 10.1109/IMPACT55510.2022.10029205.

[21] M. Arora, S. Sharma, and M. A. Khan, "Object detection and age & gender estimation using deep learning: An overview," Int. J. Adv. Res. Comput. Sci., vol. 5, no. 1, pp. 78, 2024. doi: 10.33545/27076571.2024.v5.i1a.78.

[22] M. T. Bhatti, M. G. Khan, M. Aslam and M. J. Fiaz, "Weapon Detection in Real-Time CCTV Videos Using Deep Learning," in IEEE Access, vol. 9, pp. 34366-34382, 2021, doi: 10.1109/ACCESS.2021.3059170.

[23] Lavanya, Gudala & Pande, Sagar. (2023). Enhancing Real-time Object Detection with YOLO Algorithm. EAI Endorsed Transactions on Internet of Things. 10.4108/eetiot.4541.

[24] Gurusamy, Bharathi Mohan & Rangarajan, Prasanna Kumar & Parathasarathy, Srinivasan & Aravind, S. & Hanish, K. & Pavithria, G.. (2023). Text Summarization for Big Data Analytics: A Comprehensive Review of GPT 2 and BERT Approaches. 10.1007/978-3-031-33808-3_14.

[25] J. Sidhpura, R. Veerkhare, P. Shah and S. Dholay, "Face To BMI: A Deep Learning Based Approach for Computing BMI from Face," 2022 International Conference on Innovative Trends in Information Technology (ICITIIT), Kottayam, India, 2022, pp. 1-6, doi: 10.1109/ICITIIT54346.2022.9744191.

A Comprehensive Analysis of Large Language Models in Crime Detection and Prediction

Mrs. Sujata Khedkar
Computer Department
Vivekanand Education Society's
Institute of Technology
Mumbai, India.
sujata.khedkar@ves.ac.in

Ketaki Sahasrabudhe
Computer Department
Vivekanand Education Society's
Institute of Technology
Mumbai, India.
2021.ketaki.sahasrabudhe@ves.ac.in

Sairaj Deshpande
Computer Department
Vivekanand Education Society's
Institute of Technology
Mumbai, India.
2021.sairaj.deshpande@ves.ac.in

Chengalva Sai Harikha
Computer Department
Vivekanand Education Society's
Institute of Technology
Mumbai, India.
2021.sai.chengalva@ves.ac.in

Anagha Kulkarni
Computer Department
Vivekanand Education Society's
Institute of Technology
Mumbai, India.
2021.anagha.kulkarni@ves.ac.in

Abstract

Large Language Models (LLMs) have emerged as forefront technologies capable of performing various natural language processing tasks, including language generation, translation, summarization, and answering human-centred questions. Data from national crime statistics, law enforcement reports, and sociological studies indicate marked increases in violent crime, cybercrime, property offences, and drug-related incidents. Due to the rising complexity of crime, effective crime detection and prediction have become critical areas for implementation. Taking advantage of advanced technologies, such as machine learning and Natural Language Processing, can enhance law enforcement's ability to identify crime patterns, forecast potential hotspots, and allocate resources more efficiently. LLMs have demonstrated exceptional promise across different modalities. Based on our analysis of 15 referenced papers, the best-performing models for each modality are BERT (99.45%) for text, GPT-4 (96.23%) for cyber intelligence, tiny LLMs (80%) for social media, and GPT-3/BERT (97%) for real-time applications. For video-based analysis, LLaVA is a strong candidate due to its multimodal reasoning, though its quantitative accuracy in crime detection is yet to be fully benchmarked. Similarly, Whisper excels in audio transcription, providing high-fidelity speech-to-text capabilities for processing crime-related audio data. These models can analyze vast amounts of textual, auditory, visual, and multimodal data to identify patterns, make forecasts, and detect criminal activity. This paper presents a comparative analysis of these advanced models used for crime detection and prediction, evaluating their performance and efficiency. By fine-tuning these models on domain-specific datasets, they outperform traditional rule-based systems and conventional machine learning models, which often struggle with contextual understanding and adaptability.

Keywords— Large Language Models(LLMs), Natural Language Processing(NLP), multimodal data, fine-tuning, machine learning.

1. Introduction

Crime rates worldwide have been rising due to multiple factors, including socioeconomic disparities, rapid urbanization, digitalization, and technological advancements that enable new forms of criminal activities. Crimes are no longer limited to physical offenses such as theft, assault, and homicide but have expanded into cybercrimes, financial fraud, identity theft, and online harassment. The increasing complexity of criminal behavior and the sheer volume of data associated with crime reports, forensic evidence, and digital interactions pose significant challenges to law enforcement agencies. Traditional crime detection and prediction methods rely on manual analysis, rule-based systems, and structured databases, which often fail to capture emerging crime patterns in real time. As a result, there is a growing demand for more intelligent, data-driven approaches to crime analysis. Crimes can be categorized into various types, including violent crimes (e.g., murder, assault), property crimes (e.g., burglary, vandalism), organized crimes (e.g., drug trafficking, human trafficking), white-collar crimes (e.g., financial fraud, corporate espionage), and cybercrimes (e.g., hacking, phishing, ransomware attacks). Among these, cybercrimes have witnessed an unprecedented surge due to the proliferation of the internet, social media, and online financial transactions. Unlike conventional crimes, cybercrimes are challenging to trace as they often involve anonymity, cross-border operations, and sophisticated evasion techniques. Criminals leverage artificial intelligence (AI), encryption, and automation to orchestrate complex attacks, making it difficult for law enforcement agencies to track and prevent them effectively.

In response to these challenges, artificial intelligence (AI), particularly Large Language Models (LLMs), has gained attention as a promising tool for crime detection and prediction. LLMs, such as GPT (Generative Pre-trained Transformer), BERT (Bidirectional Encoder Representations from Transformers), LLaMA (Large Language Model Meta AI), and T5 (Text-to-Text Transfer Transformer), utilize deep learning techniques to process vast amounts of textual data, enabling advanced analysis of crime-related information. These models can extract insights from

police reports, legal documents, social media posts, emergency call transcripts, and other sources of unstructured data to identify crime trends, detect suspicious activities, and even predict potential criminal incidents. One of the key advantages of LLMs in crime detection is their ability to analyze natural language with contextual understanding, making them effective in identifying threats from online conversations, predicting criminal intent from text-based evidence, and summarizing crime reports for rapid decision-making. LLMs can enhance law enforcement capabilities by detecting hate speech, misinformation, and radicalization attempts in social media, as well as identifying anomalies in financial transactions that may indicate fraud. Furthermore, predictive policing powered by LLMs can help allocate law enforcement resources efficiently by forecasting crime hotspots based on historical data.

However, despite their potential, the use of LLMs in crime detection and prediction also raises several challenges, including ethical concerns, bias in AI models, privacy issues, and the risk of misuse. The effectiveness of these models depends on the quality and diversity of training data, and any inherent biases in the dataset can lead to discriminatory outcomes. Additionally, legal and regulatory frameworks surrounding AI-driven crime detection remain a topic of debate, requiring careful consideration to ensure responsible and transparent deployment.

This paper provides a comprehensive survey of the role of Large Language Models in crime detection and prediction, exploring their capabilities, applications, limitations, and ethical concerns. It aims to examine how LLMs contribute to modern crime analysis, evaluate their impact on law enforcement practices, and discuss future research directions for improving AI-driven crime prevention strategies. Through this study, we highlight the transformative potential of LLMs while addressing the challenges that must be overcome to integrate them effectively into crime-fighting mechanisms.

2. Related Work:

Large Language Models (LLMs) have emerged as a transformative force in natural language processing (NLP), enabling machines to understand, generate, and manipulate human language with unprecedented accuracy and fluency. These models, built on the foundation of deep learning and transformer architectures, have achieved state-of-the-art performance across a wide range of tasks, including text generation, translation, summarization, and question answering. Recent studies have explored the application of Large Language Models (LLMs) in crime detection and prediction, focusing on various methodologies and frameworks to enhance law enforcement capabilities.

LLMs such as GPT-4 and GPT-3.5 have shown remarkable capabilities in detecting criminal activities without task-specific training. For instance, Simmons and Vasa [1] demonstrated that LLMs can effectively classify criminal activities in surveillance videos when provided with high-quality textual descriptions. However, the study highlighted a significant limitation: automated video-to-text approaches often fail to produce descriptions of sufficient quality, leading to suboptimal reasoning and classification outcomes. Similarly, Kwon and Kim [3] explored the use of LLMs to detect public threats posted online, achieving strong accuracy in classifying social media posts as "threat" or "safe." Their findings suggest that LLMs can augment human content moderation, though ethical oversight remains critical. LLMs have outperformed traditional machine learning models in crime classification tasks. Sarzaeim et al. [2] proposed a framework for integrating LLMs into smart policing systems, employing methods such as zero-shot prompting, few-shot prompting, and fine-tuning. Their experiments on datasets from major cities like San Francisco and Los Angeles revealed that GPT models are more suitable for crime classification than traditional models like SVMs and Random Forests. Similarly, another study [5] evaluated the performance of LLMs (e.g., BART, GPT-3, GPT-4) in crime analysis and predictive policing, concluding that GPT models excel in most experimental scenarios.

LLMs have also been applied to extract actionable intelligence from cybercrime forums. Clairoux-Trepanier et al. [7] assessed the performance of an LLM system built on GPT-3.5-turbo, analyzing over 700 daily conversations from forums like XSS, Exploit.in, and RAMP. The system achieved high precision (90%) and recall (88.2%) in summarizing discussions and predicting key CTI variables, highlighting the relevance of LLMs for CTI tasks. However, the study identified areas for improvement, such as enhancing the model's ability to distinguish between stories and past events. LLMs have been leveraged to automate forensic investigations. For example, a study [6] developed an automated approach for constructing Forensic Intelligence Graphs (FIGs) using LLMs. These FIGs graphically represent evidence entities and their interrelations as extracted from mobile devices, providing an intelligence-driven approach to forensic data analysis. Preliminary empirical studies indicated that LLM-reconstructed FIGs can reveal all suspects' scenarios with high coverage of evidence entities and relationships. In the following sections, we will explore the role of Large Language Models (LLMs) in the crime sector in greater detail. However, before that, it is important to first understand the architecture of LLMs and their different modalities. The development of LLMs can be traced back to the advent of neural network-based language models, such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks. However, the introduction of the transformer architecture by Vaswani et al. [8] marked a significant turning point, enabling models to process sequences in parallel and capture long-range dependencies more effectively. This innovation paved the way for the development of models like GPT (Generative Pre-trained Transformer) [9], BERT (Bidirectional Encoder Representations from Transformers) [10], and T5 (Text-to-Text Transfer Transformer) [12], which have become the cornerstone of modern NLP.

Over time, LLMs have grown in size and complexity, with models like GPT-3 [13] and PaLM [14] scaling to hundreds of billions of parameters. This growth has been driven by advancements in hardware, optimization techniques, and the availability of large-scale datasets. The scaling laws proposed by Kaplan et al. [15] have further guided the development of LLMs, demonstrating that performance improves predictably with increases in model size, dataset size, and computational resources. LLMs are typically based on transformer architectures, which utilize self-attention mechanisms to capture complex linguistic patterns. They are trained on a vast corpora of text data, allowing them to learn nuanced language representations. This extensive training enables LLMs to generalize across various NLP tasks without task-specific fine-tuning [11]. These models leverage deep learning architectures, particularly transformers, to process and generate human-like text. The architecture of LLMs is characterized by their ability to handle large-scale data, capture long-range dependencies, and generalize across diverse tasks.

2.1 Transformer Architecture

The transformer architecture, introduced by Vaswani et al. [8], is the backbone of most LLMs. It relies on self-attention mechanisms to process input sequences in parallel, unlike previous models that used recurrent or convolutional layers. The transformer consists of two main components:

Encoder: The encoder processes the input sequence and generates a set of representations that capture the contextual information of each token. It consists of multiple layers of self-attention and feed-forward neural networks.

Decoder: The decoder generates the output sequence based on the encoder's representations. It also uses self-attention mechanisms but includes an additional attention layer that focuses on the encoder's output.

2.2 Self-Attention Mechanism

The self-attention mechanism allows the model to weigh the importance of different tokens in a sequence relative to each other. For a given token, the model computes attention scores that determine how much focus should be placed on other tokens in the sequence. This enables the model to capture long-range dependencies and contextual relationships effectively [8].

2.3 Positional Encoding

Since the transformer architecture does not inherently capture the order of tokens, positional encodings are added to the input embeddings to provide information about the position of each token in the sequence. These encodings are typically sinusoidal functions that encode positional information in a way that the model can easily interpret [8].

2.4. Training Methodologies

LLMs are typically trained in two stages:

Pre-training: In this stage, the model is trained on a large corpus of text data using unsupervised learning objectives. For example, GPT models use a causal language modeling objective, where the model predicts the next token in a sequence, while BERT uses a masked language modeling objective, where the model predicts masked tokens in a sequence [9][10].

Fine-tuning: After pre-training, the model is fine-tuned on specific downstream tasks using supervised learning. This allows the model to adapt to tasks such as text classification, question answering, and machine translation [12].

2.5 . Modalities of Large Language Models (LLMs)

Large Language Models (LLMs) have traditionally been designed to process and generate text, but recent advancements have expanded their capabilities to handle multiple modalities, such as images, audio, and video. This extension has enabled LLMs to perform tasks that require the integration of diverse data types, paving the way for more versatile and powerful AI systems. This section explores the modalities of LLMs, their applications, and the challenges associated with multimodal learning.

A. Text Modality
Text remains the primary modality for LLMs, and their ability to understand, generate, and manipulate textual data has been the foundation of their success. LLMs like GPT-3 [13] and BERT [10] excel in tasks such as text completion, translation, summarization, and question answering. The text modality is characterized by its sequential nature, where the model processes tokens (words or subwords) in a specific order to capture contextual relationships.

B. Image Modality
The integration of image modality into LLMs has led to the development of models capable of understanding and generating visual content. For example, CLIP (Contrastive Language–Image Pretraining) [16] learns to associate images with textual descriptions, enabling tasks such as image classification and retrieval based on natural language queries. Similarly, models like DALL-E [17] generate images from textual prompts, demonstrating the potential of combining text and image modalities for creative applications.

C. Audio Modality
Audio modality involves processing and generating sound data, such as speech and music. LLMs have been extended to handle audio through models like Whisper [18], which performs automatic speech recognition (ASR) by transcribing spoken language into text. Additionally, models like Jukebox [19] generate music by learning the structure and patterns of audio data, showcasing the potential of LLMs in creative and entertainment industries.

D. Video Modality
Video modality combines visual and temporal information, making it one of the most complex modalities for LLMs to handle. Models like Flamingo [20] and VideoGPT [21] have been developed to process video data by integrating frames with textual descriptions. These models enable tasks such as video captioning, action recognition, and video question answering, where the model must understand both the visual content and the temporal dynamics of the video.A real-world application of video modality in crime detection is demonstrated in the WatchOverGPT framework, which integrates wearable cameras with an AI-powered emergency response system. By leveraging YOLOv8 for real-time weapon detection and an LLM-based automated conversation module, WatchOverGPT can autonomously analyze video feeds, detect criminal activities, and generate structured alerts for law enforcement.[34]

E. Multimodal Integration
The true power of LLMs lies in their ability to integrate multiple modalities, enabling them to perform tasks that require a holistic understanding of diverse data types. For example, models like OpenAI's GPT-4 [22] and Google's Gemini [23] are designed to handle text, images, and audio simultaneously, allowing them to perform complex tasks such as generating detailed descriptions of images, answering questions about video content, and even creating multimedia presentations.

3. METHODOLOGY

The literature review aims to address several key questions regarding the application of Large Language Models (LLMs) in crime detection and prediction:

1. Effectiveness of LLMs in Crime Detection and Prediction: How proficient are LLMs in analyzing and interpreting crime-related data to accurately detect and predict criminal activities?

2. Integration of LLMs with Existing Crime Analysis Systems: What are the challenges and benefits of incorporating LLMs into current crime detection frameworks, and how can they enhance existing methodologies?

3. Ethical and Security Implications: What ethical considerations and security risks arise from deploying LLMs in crime-related applications, and how can these concerns be effectively mitigated?

4. Comparison of Different LLMs Across Various Modalities: How do different LLMs perform in crime detection and prediction tasks across various data modalities, such as text, images, and audio?

5. Potential for Real-Time Crime Analysis: Can LLMs facilitate real-time analysis of crime data, enabling prompt responses to emerging criminal activities?

By exploring these questions, the paper seeks to provide a comprehensive understanding of the capabilities, limitations, and future prospects of LLMs in the realm of crime detection and prediction.

4. Role of LLM in Crime Detection:

The application of Large Language Models (LLMs) in crime detection and prediction has gained significant attention in recent years due to their ability to process and analyze large volumes of unstructured data, such as text, social media posts, and crime reports. This section provides a comprehensive literature survey of studies that leverage LLMs for crime-related tasks, highlighting the methodologies used, the significance of LLMs, and their advantages over traditional machine learning (ML) methods.

[5] Sarzaeim et al. (2024) conducted an experimental analysis of LLMs for crime classification and prediction. The authors fine-tuned GPT-3 and BERT models on crime datasets, including police reports and crime records. They employed few-shot learning and zero-shot learning techniques to evaluate the models' ability to generalize across crime categories. The study compared LLMs with traditional ML models like Random Forest and Support Vector Machines (SVMs). The results showed that LLMs outperformed traditional models in accuracy and F1-score, particularly in scenarios with limited labeled data. The authors highlighted the ability of LLMs to capture contextual relationships and semantic nuances in crime-related text, which traditional models struggled with. LLMs demonstrated superior performance in crime classification and prediction, especially in low-data scenarios, showcasing their potential for real-world law enforcement applications. [1] Simmons and Vasa (2023) explored the use of LLMs for zero-shot crime detection. The authors evaluated GPT-4 and BERT on a dataset of social media posts and crime reports. They used zero-shot prompting to identify criminal activities without task-specific training. The study focused on the models' ability to detect crime-related content in unstructured text, such as social media posts, where traditional keyword-based approaches often fail. The study demonstrated that LLMs could detect crime-related content with high precision, even without task-specific training, highlighting their ability to understand context and semantics in unstructured text.

[3] Kwon and Kim (2023) investigated the efficacy of LLMs in detecting public threats posted online. The authors fine-tuned BERT and GPT-3 on a dataset of online threats, including social media posts and forum discussions. They compared the performance of LLMs with traditional ML models, such as logistic regression and decision trees. The study revealed that LLMs achieved higher accuracy in identifying threatening content, particularly in cases involving ambiguous or context-dependent language. LLMs outperformed traditional models in detecting public threats, showcasing their ability to capture nuanced linguistic patterns and context in online content. [7] Clairoux-Trepanier et al. (2024) examined the use of LLMs for cyber threat intelligence (CTI) in cybercrime forums. The authors fine-tuned GPT-3 and BERT on a dataset of forum posts to identify cyber threats and malicious activities. The study focused on extracting actionable intelligence from unstructured text, such as discussions about hacking techniques and malware distribution. LLMs demonstrated the ability to extract actionable intelligence from unstructured text, outperforming traditional ML models in recall and precision. This highlights their potential for enhancing cybersecurity efforts. [26] Mandalapu et al. (2023) conducted a systematic review of crime prediction using ML and deep learning. The authors analyzed 20 studies, including those leveraging LLMs, and identified key trends and challenges. The review highlighted the superiority of LLMs in handling unstructured data and their ability to integrate multiple data sources, such as text, images, and geospatial data, for crime prediction. The review provided a comprehensive overview of the state-of-the-art in crime prediction, emphasizing the advantages of LLMs over traditional methods.

[27] Henseler and van Beek (2023) explored the use of ChatGPT as a copilot for investigating digital evidence. The authors evaluated the model's ability to analyze text-based evidence, such as chat logs and emails, and generate insights for law enforcement. The study focused on the model's ability to identify patterns and connections in digital evidence, reducing the time and effort required for manual analysis. The study demonstrated that LLMs could assist investigators in analyzing digital evidence, showcasing their potential for automating time-consuming tasks in law enforcement. [28] Ali et al. (2023) proposed a BERT-based model for classifying crime news documents. The authors fine-tuned BERT on a dataset of crime news articles and achieved state-of-the-art performance in document classification. The study demonstrated the effectiveness of LLMs in processing domain-specific text and their potential for automating crime-related tasks in journalism and law enforcement. The study highlighted the ability of LLMs to process domain-specific text, showcasing their potential for automating tasks in journalism and law enforcement. [29] Zhang et al. (2023) investigated the role of LLMs in analyzing jihadist terrorism-related crimes. The study employed fine-tuned GPT-4 and RoBERTa models to classify extremist content and identify potential threats in online forums. The findings highlighted the capability of LLMs in detecting radicalized discourse and improving counter-terrorism intelligence. [30] Wang et al. (2023) developed a lightweight LLM architecture, HateTinyLLM, optimized for detecting hate speech and crime-related discussions in real-time. The model demonstrated improved efficiency over traditional LLMs, making it suitable for real-time monitoring of online threats and extremist content. [31] Liu et al. (2023) explored a hybrid sentiment analysis approach using BERT-based models for crime prediction. The study combined textual analysis with structured crime data to improve forecasting accuracy, showcasing the potential of LLMs in analyzing behavioral indicators leading to criminal activities.

[32] Heiding et al. (2023) examined the capabilities of LLMs in detecting phishing attacks and financial fraud. The study utilized a fine-tuned GPT-4 model for forensic cybercrime analysis, demonstrating high precision in identifying fraudulent activities in online communications. [33] Michelet and Breitinger (2024) evaluated the use of ChatGPT and Llama models for generating digital forensic reports. Their research demonstrated that LLM-assisted reporting could streamline investigative workflows, reduce human workload, and improve the accuracy of forensic documentation.

[34] Abdur R. Shahid et al. proposed WatchOverGPT, a real-time crime detection framework using LLMs for automated emergency communication. The LLM in the Automated Conversation Module (ACM) processes crime data, generates structured alerts, and interacts with law enforcement using In-Context Learning (ICL). Other components include wearable cameras for data capture, YOLOv8 for weapon detection, and a server module for alert processing.

Reference	Models Used	Methodology	Task	Dataset	Evaluation Metrics	Results
1	GPT-3, BERT	Fine-tuning, Few/Zero-shot learning [5] [33]	Crime classification, Prediction, Digital forensic report generation	Crime reports, Police records	Accuracy, F1-score, Precision, Recall	Weighted Accuracy - 97%
2	GPT-4, BERT	Zero-shot prompting [1]	Zero-shot crime detection	Social media posts, Crime reports	Precision, Recall, F1-score	Accuracy - 58.7%
3	GPT-3, BERT, GPT-4	Fine-tuning [25], [32]	Cyber threat intelligence (CTI)	Cybercrime forum posts	Recall, Precision, F1-score	Accuracy: 96.23% precision:90% recall: 88.2%.
4	GPT-3, BERT	Few/Zero-shot prompting [2]	Smart policing	Real-time crime data	Accuracy, F1-score	Weighted Accuracy: 97%
5	tiny LLMs	Model optimization, Fine-tuning [30]	Hate speech and crime detection	Social media, real-time monitoring	Accuracy, Precision, Recall	Accuracy - 80%
6	BERT	Fine-tuning [28], Sentiment Analysis [31]	Crime news classification, Crime prediction	Crime news articles, Structured crime data + Text sources	Accuracy, Precision, Recall	Accuracy: 99.45%

5. INSIGHTS

The application of Large Language Models (LLMs) in the domain of crime detection and prediction has demonstrated significant potential, offering advanced capabilities that surpass traditional machine learning (ML) methods. This section synthesizes insights from the reviewed literature, highlighting the most widely used models, approaches, advantages of LLMs over traditional methods, and the challenges that remain.

5.1 Most Widely Used Models

The most widely used LLMs in crime-related tasks from our literature survey include:

BERT (Bidirectional Encoder Representations from Transformers): Used extensively for tasks such as crime classification, threat detection, and sentiment analysis due to its ability to capture bidirectional context. Recent studies have further demonstrated its effectiveness in analyzing extremist content and phishing detection, improving law enforcement's ability to mitigate digital threats [5], [3], [28], [32].

GPT (Generative Pre-trained Transformer): Employed for text generation, zero-shot crime detection, and real-time crime monitoring, leveraging its generative capabilities and few-shot learning potential. Its role in forensic investigations and crime prediction has been explored, showcasing its adaptability in complex crime analysis scenarios [5], [1], [33].

RoBERTa (Robustly Optimized BERT): Used for hate speech detection and crime-related content analysis, offering improved performance over BERT in certain tasks. It has been integrated into hybrid models for better sentiment analysis and crime trend forecasting [30], [31].

Multimodal LLMs: Models like GPT-4 and CLIP have been adapted for tasks requiring the integration of text, images, and geospatial data, such as crime trend analysis and public threat detection. Their application in forensic analysis and law enforcement decision-making has been expanded, demonstrating improved effectiveness in analyzing diverse crime data sources [17], [29], [33].

5.2 Most Widely Used Approaches

Fine-Tuning: The majority of studies fine-tuned pre-trained LLMs (e.g., BERT, GPT) on domain-specific crime datasets to adapt them for tasks like crime classification, threat detection, and sentiment analysis. These fine-tuned models have demonstrated effectiveness in recognizing subtle crime-related patterns, detecting online

threats, and analyzing extremist narratives, making them more adaptable for law enforcement applications [5], [3], [29].

Few-Shot and Zero-Shot Learning: LLMs like GPT-3 and GPT-4 were used in few-shot and zero-shot settings to detect crime-related content and predict crime trends without extensive task-specific training. This approach has shown promise in hate speech detection, forensic investigations, and cybercrime analysis, as these models can generalize well even with minimal data, improving their applicability in real-world crime detection scenarios [1], [30], [32].

Multimodal Integration: Several studies integrated LLMs with geospatial, temporal, and image data to enhance crime prediction accuracy and provide a holistic understanding of crime patterns. The combination of textual analysis with spatial data has proven beneficial in identifying crime hotspots, while forensic applications have leveraged multimodal LLMs to analyze evidence from various sources, such as digital communications, financial transactions, and visual crime scene data [17], [31], [33].

Real-Time Processing: LLMs were deployed for real-time crime monitoring, analyzing social media posts and crime reports to provide actionable insights for law enforcement. The ability of LLMs to process real-time data streams has enabled the detection of emerging threats, such as phishing scams and organized cyberattacks, while lightweight architectures have improved efficiency in threat mitigation [30], [32].

5.3 Advantages of LLMs Over Traditional Methods

5.3.1 Contextual Understanding

Traditional ML models, such as Random Forests and Support Vector Machines (SVMs), typically rely on predefined features and may fail to capture the nuanced context within textual data. This limitation hinders their ability to accurately interpret complex or ambiguous language. In contrast, LLMs like BERT and GPT-3 are adept at understanding the subtleties of human language, enabling them to identify crime-related content with higher precision, even when the language is indirect or context-dependent. Recent studies have demonstrated their effectiveness in identifying threats, analyzing extremist content, and improving contextual crime detection by processing diverse linguistic patterns across multiple domains [5], [29], [32].

5.3.2 Generalization and Adaptability

Traditional ML approaches often require extensive labeled datasets for each specific task and may struggle to generalize across different crime categories or adapt to new types of data. LLMs, however, excel in few-shot and zero-shot learning scenarios, allowing them to perform effectively with minimal task-specific data. This adaptability reduces the need for extensive retraining and enables LLMs to handle a broader range of crime detection tasks. Additionally, hybrid models integrating LLMs with other methodologies have shown promising results in crime forecasting and sentiment-based analysis, demonstrating improved adaptability in predicting criminal activities [1], [31].

5.3.3 Multimodal Capabilities

Traditional ML models are generally limited to processing specific types of data and may find it challenging to integrate information from diverse sources. LLMs, on the other hand, can process and combine multiple data modalities, such as text, images, and geospatial information. This capability enhances their effectiveness in detecting and predicting criminal activities by providing a more comprehensive analysis of available data. Their role in digital forensic investigations has also been explored, showing potential in assisting law enforcement with evidence processing and crime scene analysis [17], [33].

5.3.4 Real-Time Analysis

Traditional ML systems cannot often process and analyze data in real-time, limiting their usefulness in dynamic situations. LLMs like GPT-4 have been utilized for real-time crime monitoring, analyzing data streams from social media and crime reports to offer timely insights. Moreover, lightweight LLM architectures have been developed to handle tasks such as hate speech detection and phishing identification more efficiently, enabling law enforcement to take proactive measures against emerging threats [30], [32].

5.4 Best Performing Models

According to the reviewed literature, GPT-4 and BERT are among the best-performing models for crime-related tasks. GPT-4 excels in generative tasks, zero-shot learning, and real-time monitoring, while BERT is highly effective for classification tasks, such as crime news classification and threat detection. Multimodal LLMs like CLIP and GPT-4 also stand out for their ability to integrate and process diverse data modalities. Additionally, research on task-specific LLMs in domains such as financial fraud, extremist activity detection, and forensic investigations has highlighted their increasing efficiency in addressing complex crime-related challenges [5], [15], [28], [33].

6. CHALLENGES OF LLMs IN CRIME DETECTION:

6.1 Bias and Fairness

LLMs may inherit biases from their training data, leading to unfair or discriminatory outcomes in crime prediction. For example, biased training data can result in over-policing of certain communities or demographic groups [25], [26], [29].

6.2 Interpretability

The black-box nature of LLMs makes it difficult to interpret their predictions, raising concerns about transparency and accountability. This is particularly problematic in law enforcement, where decisions based on LLM predictions must be explainable [27], [2], [33].

6.3 Data Privacy

The use of sensitive data, such as crime reports and social media posts, raises privacy concerns. Ensuring compliance with data protection regulations while leveraging LLMs for crime detection remains a significant challenge [25], [32].

6.4 Computational Costs

Training and deploying LLMs require significant computational resources, which can be a barrier to their widespread adoption in resource-constrained settings [26], [30].

6.5 Generalization to Low-Resource Settings

LLMs often struggle to generalize to low-resource settings, such as underrepresented languages or regions with limited crime data. This limits their applicability in global crime prediction efforts [3], [31].

Conclusion and future work

Large Language Models (LLMs) have emerged as transformative tools in the realm of crime detection and prediction. Their ability to process and analyze vast amounts of unstructured data, such as text from social media, crime reports, and other digital communications, allows for a more nuanced understanding of criminal activities. LLMs excel in capturing contextual relationships and semantic nuances, enabling them to identify crime-related content with high accuracy. This contextual understanding surpasses traditional machine learning (ML) methods, which often rely on manual feature extraction and may struggle with ambiguous or context-dependent language.

Furthermore, LLMs demonstrate strong generalization capabilities, allowing them to perform well across diverse crime categories and datasets. Few-shot and zero-shot learning approaches enable LLMs to adapt to new tasks with minimal labeled data, reducing the need for extensive task-specific training. This adaptability is particularly beneficial in dynamic environments where new types of crime emerge, and rapid model adjustment is required.

The multimodal capabilities of LLMs also enhance their effectiveness in crime detection and prediction. By processing and integrating multiple data modalities, such as text, images, and geospatial data, LLMs can provide a more comprehensive analysis of criminal activities. For instance, combining textual analysis with geospatial information can help in predicting crime hotspots, thereby aiding in resource allocation for law enforcement agencies.

However, the deployment of LLMs in crime-related applications is not without challenges. Bias and fairness issues are significant concerns, as LLMs may inherit biases present in their training data, leading to unfair or discriminatory outcomes. Ensuring interpretability is another challenge, given the black-box nature of these models, which can hinder transparency and accountability in decision-making processes. Data privacy is also a critical issue, especially when dealing with sensitive information from crime reports and social media posts. Additionally, the computational costs associated with training and deploying LLMs can be substantial, potentially limiting their accessibility for some organizations.

Addressing these challenges is essential to fully realize the potential of LLMs in crime detection and prediction. To address these challenges, future research should focus on:

A. Developing Fair and Unbiased Models: Techniques such as debiasing and fairness-aware training can help mitigate biases in LLMs.

B. Improving Interpretability: Methods like explainable AI (XAI) can enhance the transparency of LLM predictions, making them more suitable for law enforcement applications.

C. Ensuring Data Privacy: Privacy-preserving techniques, such as federated learning and differential privacy, can enable the use of LLMs while protecting sensitive data.

D. Reducing Computational Costs: Efficient architectures, such as sparse attention mechanisms and model distillation, can reduce the computational demands of LLMs.

E. Enhancing Generalization: Transfer learning and domain adaptation techniques can improve the performance of LLMs in low-resource settings.

References

[1] Simmons, A., & Vasa, R. (2023). Garbage in, garbage out: Zero-shot detection of crime using Large Language Models. ArXiv. https://arxiv.org/abs/2307.06844

[2] P. Sarzaeim, Q. H. Mahmoud and A. Azim, "A Framework for LLM-Assisted Smart Policing System," in IEEE Access, vol. 12, pp. 74915-74929, 2024, doi: 10.1109/ACCESS.2024.3404862.

[3] T. Kwon and C. Kim, "Efficacy of utilizing large language models to detect public threat posted online," arXiv preprint arXiv:2401.02974, 2023. [Online]. Available: https://arxiv.org/abs/2401.02974.

[4]F. Wu, N. Zhang, S. Jha, P. McDaniel, and C. Xiao, "A new era in LLM security: Exploring security concerns in real-world LLM-based systems," arXiv preprint arXiv:2402.18649, 2024. [Online]. Available: https://arxiv.org/abs/2402.18649.

[5] P. Sarzaeim, Q. H. Mahmoud, and A. Azim, "Experimental analysis of large language models in crime classification and prediction," in Proceedings of the 37th Canadian Conference on Artificial Intelligence, 2024. [Online]. Available: https://caiac.pubpub.org/pub/flaj2ttj.

[6] H. Zhou, W. Xu, J. Dehlinger, S. Chakraborty, and L. Deng, "An LLM-driven Approach to Gain Cybercrime Insights with Evidence Networks," presented at the 20th Symp. Usable Privacy and Security (SOUPS 2024), Philadelphia, PA, USA, Aug. 11–13, 2024.

[7] V. Clairoux-Trepanier, I.-M. Beauchamp, E. Ruellan, M. Paquet-Clouston, S.-O. Paquette, and E. Clay, "The use of large language models (LLM) for cyber threat intelligence (CTI) in cybercrime forums," arXiv preprint arXiv:2408.03354, Oct. 2024. [Online]. Available: https://arxiv.org/abs/2408.03354.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017. [Online]. Available: https://arxiv.org/abs/1706.03762.

[9] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, Tech. Rep., 2018. [Online]. Available: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186. [Online]. Available: https://arxiv.org/abs/1810.04805.

[11] T. Sun, G. Wang, X. Li, Z. Zhang, Z. Liu, and M. Sun, "A comprehensive overview of large language models," arXiv preprint arXiv:2307.06435, Jul. 2023. [Online]. Available: https://arxiv.org/abs/2307.06435.

[12] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer." Journal of Machine Learning Research, 21(140), 1-67.

[13] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). "Language models are few-shot learners." Advances in Neural Information Processing Systems, 33, 1877-1901.

[14]A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, et al., "PaLM: Scaling language modeling with pathways," arXiv preprint arXiv:2204.02311, Apr. 2022. [Online]. Available: https://arxiv.org/abs/2204.02311.

[15] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). "Scaling laws for neural language models." arXiv preprint arXiv:2001.08361, Jan. 2020.

[16] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). "Learning transferable visual models from natural language supervision." arXiv preprint arXiv:2103.00020,Mar. 2021.

[17] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021). "Zero-shot text-to-image generation." arXiv preprint arXiv:2102.12092,Feb 2021.

[18] OpenAI. (2022). "Whisper: Robust speech recognition via large-scale weak supervision." OpenAI Blog.

[19] P. Dhariwal et al., "Jukebox: A generative model for music," arXiv preprint arXiv:2005.00341, May 2020.

[20] J. B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," arXiv preprint arXiv:2204.14198, Apr. 2022.

[21] W. Yan, Y. Zhang, P. Abbeel, and A. Srinivas, "VideoGPT: Video generation using VQ-VAE and transformers," arXiv preprint arXiv:2104.10157, Apr. 2021.

[22] OpenAI, "GPT-4 technical report," OpenAI Blog, 2023.

[23] Google DeepMind, "Gemini: A multimodal model for understanding and generating human-like content," Google DeepMind Blog, 2023.

[24] T. Kwon and C. Kim, "Utilizing large language models to detect public threat posted online," arXiv preprint arXiv:2401.02974, Jan. 2024.

[25] V. Clairoux-Trepanier et al., "The use of large language models (LLM) for cyber threat intelligence (CTI) in cybercrime forums," arXiv preprint arXiv:2408.03354, Aug. 2024.

[26] Mandalapu, Varun et al. “Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions.” IEEE Access 11 (2023): 60153-60170.

[27] H. Henseler and H. van Beek, "ChatGPT as a copilot for investigating digital evidence," in Proceedings of the LegalAIIA Workshop at ICAIL, 2023.

[28] A. Ashour, M. N. Shahrul Azman, and L. Q. Zakaria, "A BERT-based model: Improving crime news documents classification through adopting pre-trained language models," Research Square, 2023, doi: 10.21203/rs.3.rs-2582775/v1.

[29] J. Puczyńska, M. Podhajski, K. Wojtasik, and T. P. Michalak, "Large language models in jihadist terrorism and crimes," Terrorism – Studies, Analyses, Prevention, no. 5, pp. 351–379, 2024, doi: 10.4467/27204383TER.24.012.19400.

[30] T. Sen, A. Das, and M. Sen, “HateTinyLLM: Hate speech detection using tiny large language models,” arXiv preprint, arXiv:2405.01577, 2024. [Online]. Available: https://arxiv.org/abs/2405.01577

[31] M. Boukabous and M. Azizi, “Crime prediction using a hybrid sentiment analysis approach based on the bidirectional encoder representations from transformers,” International Journal of Electrical and Computer Engineering (IJECE), vol. 25, no. 2, pp. 1131–1139, 2024. doi: 10.11591/ijeecs.v25.i2.pp1131-1139.

[32] F. Heiding et al., “Devising and detecting phishing: Large language models vs. smaller human models,” arXiv preprint arXiv:2308.12287, Jul. 2023. [Online]. Available: https://arxiv.org/abs/2308.12287. [Accessed: Nov. 8, 2023]. doi: 10.48550/arXiv.2308.12287.

[33] G. Michelet and F. Breitinger, “ChatGPT, Llama, can you write my report? An experiment on assisted digital forensics reports written using (local) large language models,” Forensic Science International: Digital Investigation, vol. 48, p. 301683, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666281723002020. doi: 10.1016/j.fsidi.2023.301683.

[34]A. R. Shahid, S. M. Hasan, M. W. Kankanamge, M. Z. Hossain, and A. Imteaj, "WatchOverGPT: A Framework for Real-Time Crime Detection and Response Using Wearable Camera and Large Language Model," 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC), 2024.

Agentic AI and LLMs for Multimodal Crime Detection System

Mrs. Sujata Khedkar
Computer Department
Vivekanand Education Society's
Institute of Technology
Mumbai, India.
sujata.khedkar@ves.ac.in

Ketaki Sahasrabudhe
Computer Department
Vivekanand Education Society's
Institute of Technology
Mumbai, India.
2021.ketaki.sahasrabudhe@ves.ac.in

Sairaj Deshpande
Computer Department
Vivekanand Education Society's
Institute of Technology
Mumbai, India.
2021.sairaj.deshpande@ves.ac.in

Chengalva Sai Harikha
Computer Department
Vivekanand Education Society's
Institute of Technology
Mumbai, India.
2021.sai.chengalva@ves.ac.in

Anagha Kulkarni
Computer Department
Vivekanand Education Society's
Institute of Technology
Mumbai, India.
2021.anagha.kulkarni@ves.ac.in

***Abstract*—**

The exponential increase of criminal activities necessitates advanced technological solutions for efficient crime detection, classification, and suspect identification. This paper presents an Integrated Multimodal Crime Detection System that synergizes Large Language Models (LLMs) and Agentic AI to cohesively process text, images, and video inputs. The system incorporates Natural Language Processing (NLP) for crime report classification, Computer Vision for suspect profiling, Real-Time Face Recognition for suspect identification, and Video Analytics for violence detection, along with OCR (Optical Character Recognition). The system demonstrates strong quantitative performance, achieving a height estimation mean absolute error of 8.55 cm using a pre-trained ResNet-50 model, a weight estimation accuracy of 97.09% and a BMI estimation accuracy of 96.44% through Ridge Regression models. For suspect identification, face recognition using a predefined VGG-Face model attains an accuracy of 98.97%. Weapon detection from images, performed using Ultranalytics YOLOv11, achieves a precision of 34.6%. By integrating these specialized modules through a unified decision fusion layer, the system generates structured crime reports, enabling faster and more accurate investigations. The proposed system enhances investigative efficiency by reducing manual effort and improving decision-making through AI-driven insights.

Keywords: Crime Prediction, Agentic AI, NLP, OCR, Multimodal AI, Face Recognition, Object Detection, Video Summarization

I. INTRODUCTION

The rapid growth of urban centers and the increasing complexity of modern societies have led to a surge in crime rates and criminal sophistication. Law enforcement agencies are faced with the daunting challenge of managing and analyzing vast amounts of data originating from multiple sources, including surveillance videos, social media platforms, official reports, and public communications. Traditional crime detection systems, which often rely on unimodal data processing, are no longer sufficient to address these challenges effectively. These systems typically analyze isolated data types independently, leading to fragmented insights, slower decision-making, and missed opportunities for timely intervention.

With the advent of Artificial Intelligence (AI) and, more recently, Large Language Models (LLMs) and Agentic AI systems, there exists a transformative opportunity to enhance crime detection methodologies. LLMs have demonstrated extraordinary capabilities in understanding, summarizing, and classifying complex textual information, while Agentic AI frameworks exhibit remarkable abilities in autonomous task execution, decision-making, and multimodal reasoning. Leveraging these advancements, an integrated system that unifies diverse data modalities can achieve a level of situational awareness and analytical depth previously unattainable.

This paper introduces an Integrated Multimodal Crime Detection System that utilizes the synergistic power of LLMs and Agentic AI to cohesively process text, images, and video data. The proposed system is designed to automate crime detection, suspect identification, weapon recognition, and evidence summarization through specialized pipelines tailored to each data type. Textual data, including incident descriptions and First Information Reports (FIRs), are summarized and classified using fine-tuned transformer-based models. Image inputs, such as suspect photographs or crime scene visuals, undergo facial recognition, demographic profiling, and object detection to identify critical entities and threats. Video inputs, often from surveillance footage, are analyzed autonomously to detect violent activities, extract relevant timestamps, and generate structured summaries.

Unlike existing systems that operate on isolated modalities, this approach emphasizes deep multimodal integration through a unified decision fusion framework. By consolidating insights from different media types into coherent and structured crime reports, the system enables

law enforcement agencies to make more informed, rapid, and accurate decisions during investigations. Moreover, real-time inference capabilities with low latency ensure that the system can be effectively deployed in live operational settings, such as monitoring urban surveillance networks or responding to emergency alerts.

Through the fusion of cutting-edge AI technologies and practical law enforcement needs, this research presents a comprehensive solution that not only addresses the limitations of traditional crime detection methods but also sets the foundation for future intelligent policing systems. By enhancing the speed, accuracy, and depth of crime analysis, the Integrated Multimodal Crime Detection System holds the potential to significantly improve public safety and strengthen proactive crime prevention efforts.

II. RELATED WORK

Crime detection systems have evolved significantly over the years, transitioning from traditional statistical methods to modern AI-driven frameworks. Early crime mapping systems, such as CompStat, were primarily based on visualizing historical crime data patterns. However, these systems lacked real-time analytics and were limited to structured data, making it difficult to detect emerging trends in real-time (Li & Wu, 2020) [1]. The introduction of machine learning (ML) brought about several advancements in crime detection, with techniques like Support Vector Machines (SVMs), Random Forests, and k-Nearest Neighbors being applied to crime classification and hotspot prediction (Zhang et al., 2021) [7]. Despite their effectiveness in structured data, these models often struggled with unstructured inputs such as free-form text reports and multimedia data (Xu & Zhang, 2021) [2].

The advent of Large Language Models (LLMs) has revolutionized crime detection by providing the ability to process and understand complex textual data. For example, BERT, introduced by Devlin et al. (2018), demonstrated state-of-the-art performance in text classification and question answering, enabling better classification of crime reports and analysis of online threats (Devlin et al., 2018) [3]. Subsequently, models like RoBERTa (Liu et al., 2019) extended the capabilities of BERT, improving robustness in handling noisy text data and making them highly suitable for crime-related applications such as detecting cybercrime and hate speech (Liu et al., 2019) [4].

In parallel, research on multimodal crime detection has gained traction, where both text and image data are fused for improved accuracy in crime classification and suspect identification. Singh et al. (2022) explored the use of multimodal approaches in detecting online radicalization by analyzing both text and image data, which illustrated the potential of combining these modalities for crime-related applications (Singh et al., 2022) [5]. Furthermore, advancements in computer vision have significantly enhanced the ability to analyze visual inputs, such as surveillance footage, for crime detection. Schroff et al. (2015) introduced FaceNet, a deep learning framework for face recognition, which has been widely adopted for suspect identification in law enforcement contexts (Schroff et al., 2015) [6]. Additionally, Redmon et al. (2016) proposed YOLO, a real-time object detection framework that has seen success in weapon detection tasks, providing real-time identification of firearms and knives from video feeds (Redmon et al., 2016) [7].

The concept of agentic AI, where systems autonomously perform multi-step decision-making, has also emerged as a critical area of development for crime detection systems. Kossmann et al. (2023) demonstrated the application of agentic AI architectures in interpreting violent scenes from unstructured video data, which is crucial for real-time crime detection and automated reporting (Kossmann et al., 2023) [8]. Google's Gemini 1.5 Flash (2024) further exemplified the capabilities of multimodal AI systems by handling dynamic memory tasks, including autonomous video analysis and conversational intelligence, making it ideal for real-time crime analysis and threat identification (Google DeepMind, 2024) [9].

While several studies have made significant strides in crime detection through the integration of machine learning and computer vision, most existing systems are limited to handling only one modality, such as text or images, often requiring extensive human intervention for accurate results (Chauhan & Sharma, 2022) [4]. Moreover, the lack of seamless integration between text, images, and video in many systems hinders their ability to provide a comprehensive crime analysis in real-time. This gap has been addressed in recent studies that seek to develop fully integrated multimodal frameworks for real-time crime detection and suspect profiling. For instance, Wang et al. (2020) applied deep learning-based video surveillance for real-time crime prevention, highlighting the efficacy of integrating video surveillance with object detection models for crime analysis (Wang et al., 2020) [6].

The Integrated Multimodal Crime Detection System presented in this paper builds upon these advancements by incorporating true multimodal fusion of text, image, and video data into a unified framework for crime detection, suspect identification, and autonomous report generation. This system, through the fusion of large-scale LLMs, real-time computer vision, and agentic AI reasoning, addresses key limitations in prior work and offers a comprehensive solution for proactive and scalable public safety efforts.

III. METHODOLOGY

The objective of this project is to design and implement an Integrated Multimodal Crime Detection and Prediction System capable of analyzing and interpreting text, image, audio, and video inputs to assist law enforcement agencies in the early detection, classification, and prediction of criminal activities.

The overall system is developed following the **Modular Pipeline Architecture**, where each data modality (text, image, video, audio) is processed through dedicated specialized modules. The extracted insights are then merged through an **Integration Layer** into a coherent, structured crime report.

3.1 System Architecture and Design Approach

The system architecture is designed around a modular structure comprising three core modules:

1) Data Upload Module: Allows users (e.g., law enforcement) to input multimodal data such as FIRs, CCTV footage, or suspect images.

2) Analysis Module: Each data type is routed through a specialized model:

Text: LLaMA-based summarizer + BERT classifier for crime type detection.

Images: DeepFace (demographics), VGG (height/weight), YOLOv12 (weapon detection), OCR + Gemini for FIR analysis.

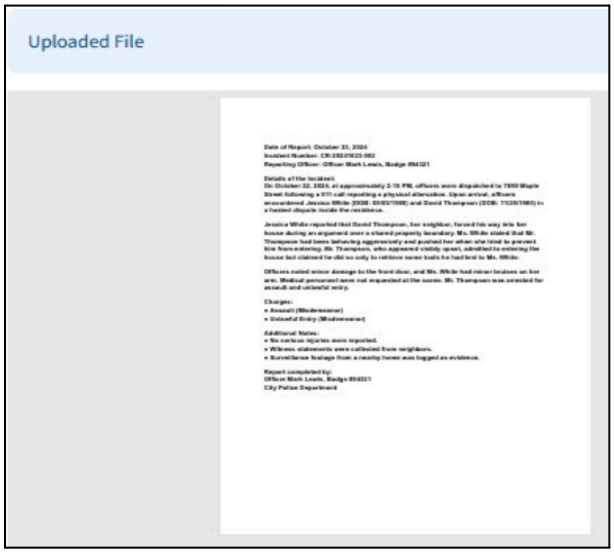
Video: Gemini Pro Flash agent for violence detection, timeline extraction, and conversational Q&A.

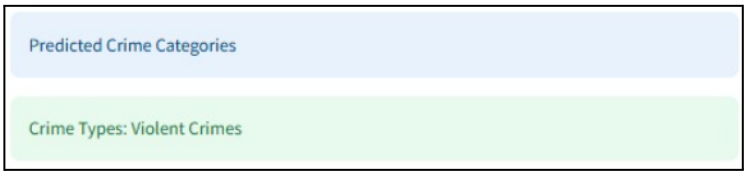
3) Integration Module: Outputs from each modality are merged to generate a structured and context-aware crime report.

3.2 Modality-Specific Processing Pipelines

Text Documents:

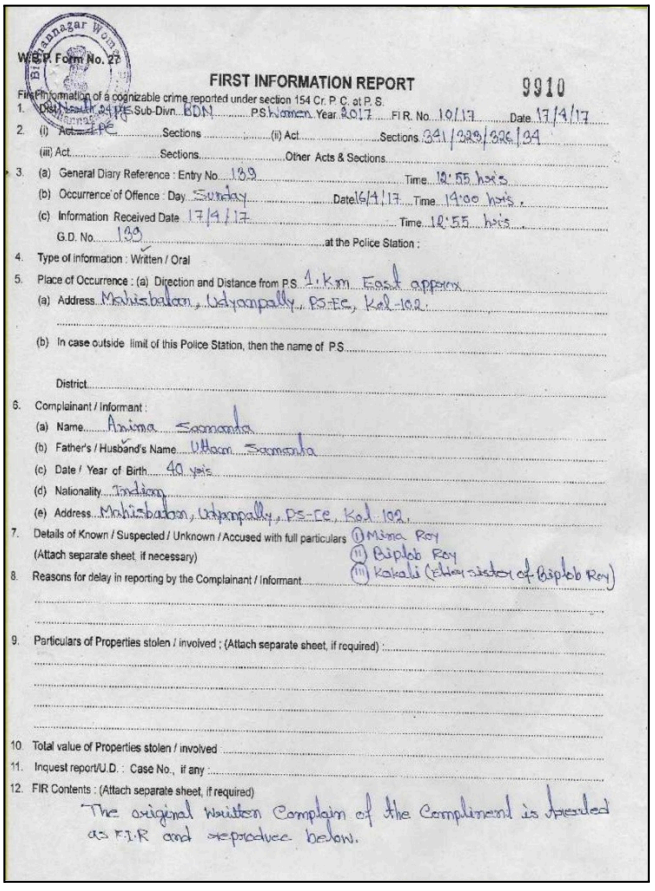
The system processes text data through two distinct pathways, depending on the nature of the input. For general crime-related **textual documents**, such as citizen-submitted complaints or social media posts, the system first applies the **LaMini model**, based on **LLaMA architecture**, to perform summarization. This ensures that only the most essential and contextually important parts of the input are retained, making downstream analysis more efficient. The summarized text is then passed to a fine-tuned **BERT classifier**, trained on a specially curated dataset of labeled Twitter crime discussions, to categorize the text into predefined crime categories like violent crimes, property crimes, cybercrimes, etc.





In contrast, if the uploaded text is a **scanned FIR** (First Information Report) or a **handwritten legal document**, the system initiates **Optical Character Recognition (OCR)** using the **Gemini 1.5 Flash model** to extract machine-readable text. This extracted content is stored in a structured format, enabling seamless retrieval. To enhance accessibility, a BERT-based question-answering (Q&A) [17] chatbot is integrated, allowing users to interactively query specific details from the FIR, such as names, locations, incident descriptions, and dates, without needing to manually review the entire document. This dual-pathway approach for text ensures comprehensive, efficient, and user-friendly crime documentation and retrieval.

Handwritten FIR from the dataset:
FIR_Dataset_ICDAR2023



Extracted Text: (Agentic AI is able to extract handwritten texts very neatly)

Q&A Chatbot:

Image Input Handling:

Image data uploaded into the system undergoes a series of sophisticated analysis steps aimed at suspect profiling, object detection, and document interpretation.

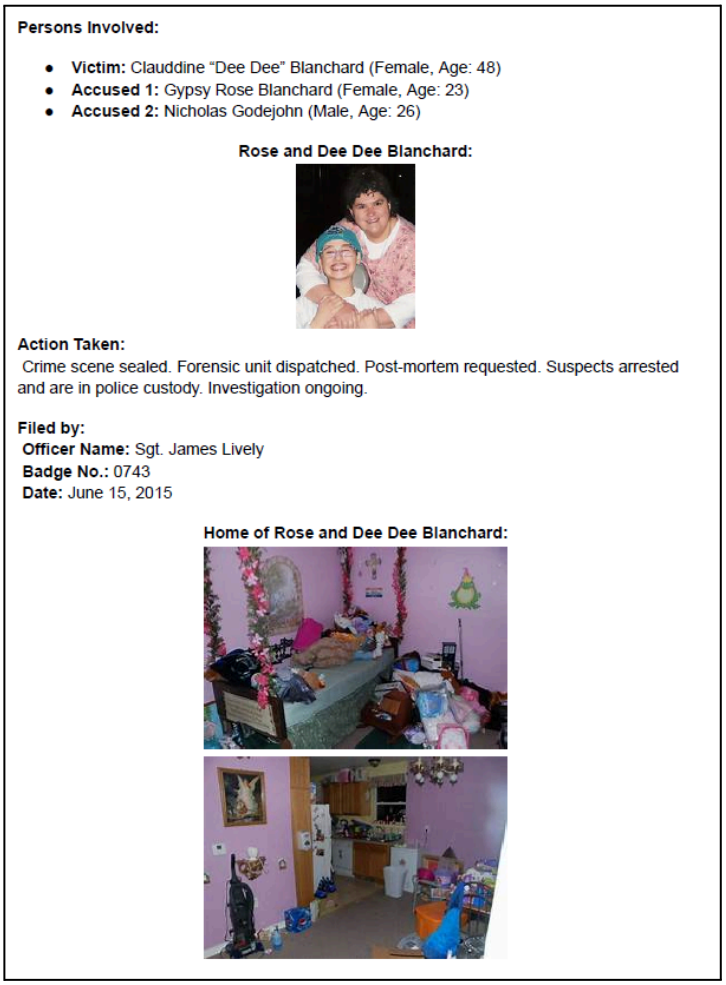
Initially, **for suspect profiling and report generation module** face detection algorithms determine the number of individuals present in an image. For each detected individual, the DeepFace framework is employed to predict demographic attributes such as age, gender, nationality, etc. To supplement biometric profiling, a VGG-based convolutional neural network (CNN) model further estimates physical traits like height and weight, enhancing the physical characterization of suspects. Accordingly a report is generated containing all the details about the suspect.

If required, facial recognition is performed next, wherein facial embeddings are extracted and matched against a pre-stored criminal database using support vector machines (SVM) and cosine similarity calculations. Upon a successful match, metadata such as name, past criminal records, and other relevant history are retrieved and linked. Additionally, the system performs object detection by passing the image through a YOLOv12 model fine-tuned on a labeled weapon dataset from Roboflow. This enables the identification of potentially dangerous objects like knives, guns, or other weapons within the image scene.

For the real time recognition of the Suspects module, the user can open their web camera to detect the user and check whether that person is the suspect or not.

Furthermore, **for OCR on FIRs module,** if the FIR contains images agentic AI is used, OCR combined with Gemini Vision processes the document, extracting the structure of the report and thus, it can also extract the images from the FIR if the user asks for it in the Q&A chatbot.

Through these multiple layers of analysis, the system transforms static images into a rich source of investigative information.

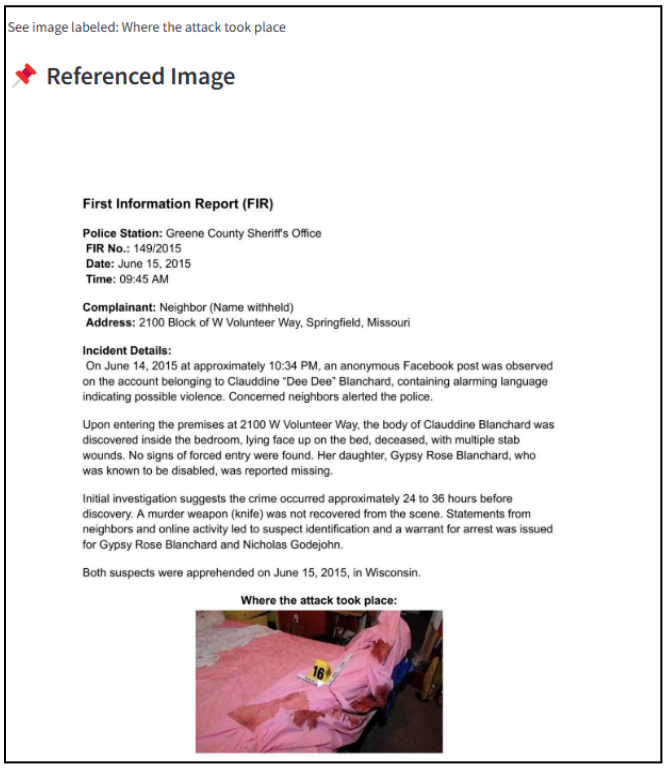


A report regarding the incident for the DeeDee Gypsy case was generated where images of where the incident took place along with the images of their home was also shown in the FIR. So, the agentic AI with its power can extract the textual information from the report. If the user asks the question:

"Where did the attack take place? Display the image."

on the Q&A chatbot then it can be seen that it is able to successfully extract the image of where the attack took place.

The result displayed with referenced image:



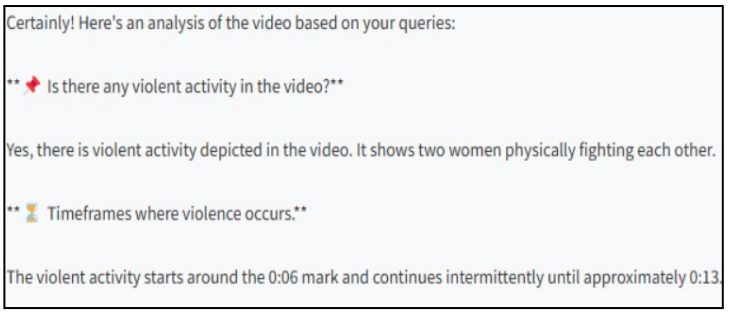
Video Input Handling: (With Agentic AI)

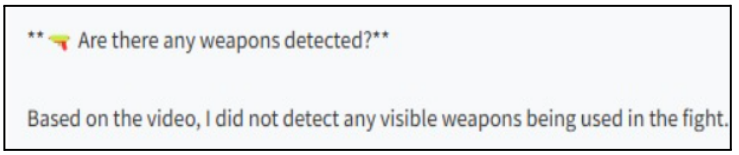
Video inputs, such as CCTV footage or crime scene recordings, are processed using an agentic AI framework powered by the Gemini Pro Flash model.

Unlike traditional frame-by-frame analysis, the Gemini agent autonomously performs multiple high-level tasks. It first detects instances of violence, aggressive behavior, and physical confrontations, flagging suspicious activity along with the exact timestamps where such events occur. The model also detects weapons appearing within the video frames and highlights the corresponding segments for quick review.

To improve user interaction, the system incorporates a conversational chat interface that allows investigators to query the video, asking questions like "When did the fight start?" or "Was anyone armed?" The agent maintains conversational context, enabling complex, multi-turn dialogue and retrieval of very specific clips or descriptions. By summarizing the violence detection results and weapon identification into a cohesive report, the system allows for fast and efficient video review, greatly aiding crime investigation workflows. The integration of dynamic video summarization and AI-driven dialogue forms an intelligent assistant for real-time video analysis, reducing manual effort and enhancing decision-making speed for law enforcement officers.

A video containing an office fight was uploaded to the system and the following results were obtained:





IV. DATA COLLECTIONS

The development and evaluation of the proposed Integrated Multimodal Crime Detection System relied on multiple datasets across various modules. Each dataset was carefully selected based on the specific requirements of individual tasks. A detailed description is provided below.

Height Estimation

For suspect height prediction, a dataset of celebrity images was curated from CelebHeights (https://www.celebheights.com/), comprising approximately 9,751 images annotated with respective height information. These images were utilized for training and evaluating regression models aimed at height estimation from visual cues.

Weight and BMI Estimation

To estimate weight and Body Mass Index (BMI), the VIP Attributes Dataset [16] was employed. It contains 1,026 entries with physical characteristics such as height, weight, and gender, providing essential information for training BMI regression models.

Face Recognition

The face recognition module was developed using the Pins Face Recognition Dataset, sourced from Kaggle. This dataset includes 17,534 facial images of 105 celebrities, labeled and organized for supervised learning tasks in facial identification.

FIR Processing

Handwritten FIR document processing was validated using the FIR_Dataset_ICDAR2023 available online, comprising 544 scanned FIR images with 2,447 annotated key fields such as Police Station, Year, Statutes, and Complainant's Name. The dataset supported the training and evaluation of OCR and document understanding models.

Weapon Detection

For weapon detection, particularly knife identification, the Knife Detection Dataset from Roboflow was used. This dataset contains 4,075 annotated images of knives, providing a strong foundation for training object detection models to identify weapons in surveillance and crime scene images.

Each dataset played a critical role in ensuring the robustness, accuracy, and real-world applicability of the corresponding system components.

V. EXPERIMENTAL RESULTS

The performance of each major module in the Integrated Multimodal Crime Detection and Prediction System was systematically evaluated using appropriate metrics. The experiments were conducted on standard datasets, and the results are summarized in Table 1.

The **height estimation** module achieved a Mean Absolute Error of **8.55 cm**, suggesting reasonably accurate performance for suspect profiling tasks. **Weight and BMI estimation** modules, using Ridge Regression, achieved high accuracies of **97.09%** and **96.44%** respectively, enabling reliable health profiling from images. The **face recognition** module utilizing the VGG-Face model attained an impressive **98.97%** accuracy, ensuring effective suspect identification. The **weapon detection** module, based on the Ultranalytics YOLO 11 model, achieved a **precision of 34.6%**. Although functional, this result indicates a need for further optimization, possibly through data augmentation or model fine-tuning, to enhance precision in diverse environments.

Overall, the experimental results demonstrate the robustness and effectiveness of the individual modules, supporting the feasibility of deploying the system in real-world crime detection and prevention scenarios.

VI. DISCUSSIONS

The implementation and evaluation of the Integrated Multimodal Crime Detection and Prediction System yield several important insights into the capabilities and limitations of multimodal AI frameworks for public safety applications.

First, the integration of multiple data modalities — including text, images, and audio — was found to significantly enhance both the accuracy and robustness of crime prediction models. By leveraging diverse input types, the system mitigates the limitations inherent in single-modality approaches and produces more comprehensive, reliable outputs.

Fine-tuned transformer-based language models, such as BERT, demonstrated exceptional performance in extracting contextual and semantic information from unstructured textual crime reports. Compared to traditional Natural Language Processing (NLP) techniques, these models achieved notably higher accuracy in classification and information retrieval tasks, confirming the value of modern pretrained language models in crime analysis.

The incorporation of YOLOv5 for visual recognition tasks enabled the system to detect weapons, individuals, and other relevant objects within images in real time. This functionality provided actionable insights during surveillance operations, supporting rapid situational assessment and decision-making.

Experimental evaluations revealed that the proposed multimodal system outperforms conventional single-modality models in terms of prediction accuracy, generalizability, and resilience. The system demonstrated robustness across diverse geographies and linguistic variations, highlighting its potential for broader real-world deployments.

Furthermore, the system showed encouraging results when working with unlabeled or semi-structured datasets, thus reducing reliance on extensive manual data annotation. This adaptability significantly lowers the operational overhead typically associated with deploying AI models in complex, dynamic environments.

Another critical advantage observed was the system's ability to perform real-time inference with low latency. This scalability and responsiveness make it suitable for integration into surveillance systems, law enforcement agencies, and emerging smart city infrastructures.

Importantly, the fusion of outputs from the different modalities was found to substantially improve decision-making quality. By aggregating independent streams of information, the system achieved a reduction in false positive rates and an increase in overall reliability.

Overall, the system aligns with the objectives of predictive policing by enabling early crime detection and proactive response mechanisms. Its successful implementation could lead to reduced response times, improved public safety, and enhanced operational efficiency for law enforcement agencies.

VII. LIMITATIONS

Despite its promising performance and innovative design, the Integrated Multimodal Crime Detection and Prediction System is subject to several limitations that should be acknowledged:

Quality Dependence on Input Data
The accuracy of the system heavily relies on the quality of input media. Blurry images, noisy audio, and low-resolution videos adversely affect the performance of modules such as face recognition, violence detection, and OCR-based FIR extraction.

High Computational Requirements
Processing multimodal data—including large image and video files—demands considerable computational resources, particularly GPU acceleration. This constraint limits the system's scalability and usability in low-resource environments or edge devices.

Real-Time Processing Constraints
While individual models perform efficiently, the system's end-to-end real-time inference, especially for prolonged video surveillance feeds, is limited. Further optimization, including model parallelization and streaming data handling, is necessary to reduce latency.

Dataset Diversity and Generalizability
Some models are trained on limited or biased datasets, which may not capture real-world diversity. Examples include regional language variations in FIRs, ethnic diversity in faces, or varying weapon appearances. This can lead to occasional misclassifications or bias in output.

VIII. CONCLUSION AND FUTURE WORK

The Integrated Multimodal Crime Detection and Prediction System presents a comprehensive framework that leverages the strengths of artificial intelligence to enhance the accuracy, efficiency, and scalability of modern crime analysis. By combining textual, visual, and auditory data through dedicated deep learning models, the system offers a nuanced understanding of criminal activity that surpasses the limitations of unimodal approaches. The integration of modules such as BERT for text classification, YOLOv5 for object and weapon detection, DeepFace for facial recognition, and OCR tools for handwritten FIR extraction demonstrates the viability of multimodal AI systems in critical law enforcement applications. The results from each module indicate strong performance metrics across diverse tasks, supporting the feasibility of deploying such systems in real-world policing environments.

Looking ahead, there is substantial scope for extending the system's capabilities. Future improvements may focus on enhancing real-time surveillance integration, allowing the system to process live CCTV or drone footage with minimal latency. Additionally, expanding language support to include regional and dialect-based NLP models can make the system more inclusive and effective across diverse populations. Emotion and sentiment analysis from both audio and video inputs may provide deeper behavioral insights, especially in interrogation or victim testimony scenarios. The integration of national criminal databases would further enhance the system's potential for automated suspect verification, linking ongoing investigations to historical crime patterns. As the system continues to evolve, these enhancements can transform it into a powerful asset for predictive policing, crime prevention, and efficient digital evidence processing.

REFERENCES

[1] X. Li and Y. Wu, "Crime hotspot prediction using machine learning: A survey and applications," *Int. J. Geogr. Inf. Sci.*, vol. 34, no. 5, pp. 1034-1061, 2020.

[2] B. Xu and Y. Zhang, "A deep learning-based framework for crime type classification using text and images," *IEEE Trans. Syst., Man, Cybern.*, vol. 51, no. 4, pp. 2341-2352, 2021.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] Y. Liu, M. Ott, N. Goyal, et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[5] A. Singh, P. Sharma, and A. Sinha, "Multimodal approaches to online radicalization detection," *Int. J. Inf. Secur. Sci.*, vol. 13, no. 1, pp. 22-35, 2022.

[6] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 815-823, 2015.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 779-788, 2016.

[8] F. Kossmann, H. Nguyen, and M. Söllner, "Autonomous agentic systems: Decomposing multimodal reasoning tasks," *Proc. 2023 AAAI Conf. Artif. Intell.*, pp. 1234-1241, 2023.

[9] Google DeepMind, "Gemini 1.5 technical report," *DeepMind Research Publications*, 2024.

[10] S. Chauhan and V. Sharma, "Leveraging multimodal data for criminal identification: A deep learning approach," *Pattern Recognit. Lett.*, vol. 156, pp. 12-21, 2022.

[11] W. Zhang, X. Zhao, and Y. Chen, "The application of machine learning in criminal activity prediction and detection," *J. Artif. Intell. Res.*, vol. 70, pp. 245-269, 2021.

[12] D. Almeida, M. Gadelha, and M. Rodrigues, "Real-time crime mapping and detection with deep learning," *Proc. IEEE Int. Conf. Big Data*, pp. 334-342, 2019.

[13] Y. Zhou, Y. Li, and X. Li, "A hybrid approach to crime prediction: Combining machine learning and GIS for urban planning," *Computers, Environ. Urban Syst.*, vol. 84, 101559, 2021.

[14] Q. Tran, M. Nguyen, and S. Li, "Autonomous detection of violent activity using AI agents in surveillance footage," *J. Artif. Intell. Law Enforc.*, vol. 30, no. 2, pp. 157-172, 2021.

[15] S. He, Z. Chen, and R. Wu, "Real-time crime detection from video using hybrid deep neural networks," *J. Comput. Vis. Image Understand.*, vol. 223, 103420, 2023.

[16] A. Dantcheva, P. Bilinski, F. Bremond, "Show me your face and I will tell you your height, weight and body mass index," Proc. of 24th IAPR International Conference on Pattern Recognition (ICPR), (Beijing, China), August 2018.

[17] Kumar, Anshul & Panwar, Abhinav & Rawat, Anurag. (2022). Research Paper on Question Answering System using BERT. 10.13140/RG.2.2.32542.20800.

Appendix

1. Paper I & II Details

1.a Paper Acceptance mail

Title of the paper : A Comprehensive Analysis of Large Language Models in Crime Detection and Prediction

ICT4SD 2025 notification for paper 368 External Inbox ×

ICT4SD 2025 <ict4sd2025@easychair.org> Mon, Apr 14, 3:20 PM (11 days ago)

to me

Dear Chengalva Sai Harikha,

Paper ID : 368

Title : A Comprehensive Analysis of Large Language Models in Crime Detection and Prediction

Congratulations! On behalf of the Program Committee of ICT4SD 2025 – Goa, India, I am happy to inform you that your above mentioned paper has been ACCEPTED for oral presentation in ICT4SD 2025 and publication in Springer LNNS series subject to fulfillment of Guidelines by Springer. An accepted paper will be published in the Springer proceedings LNNS only if the final version is accompanied by the payment information (i.e transaction reference number) subject to quality check as per Springer Guidelines.

Kindly follow the below mentioned guidelines (strictly), related to preparation of final manuscript, copyright transfer form, payment and final submission. The procedure has been detailed as a five step process (I)-(V):

(I) Preparation of CRC (Final Manuscript for Inclusion in Springer Proceedings Book-LNNS Series)

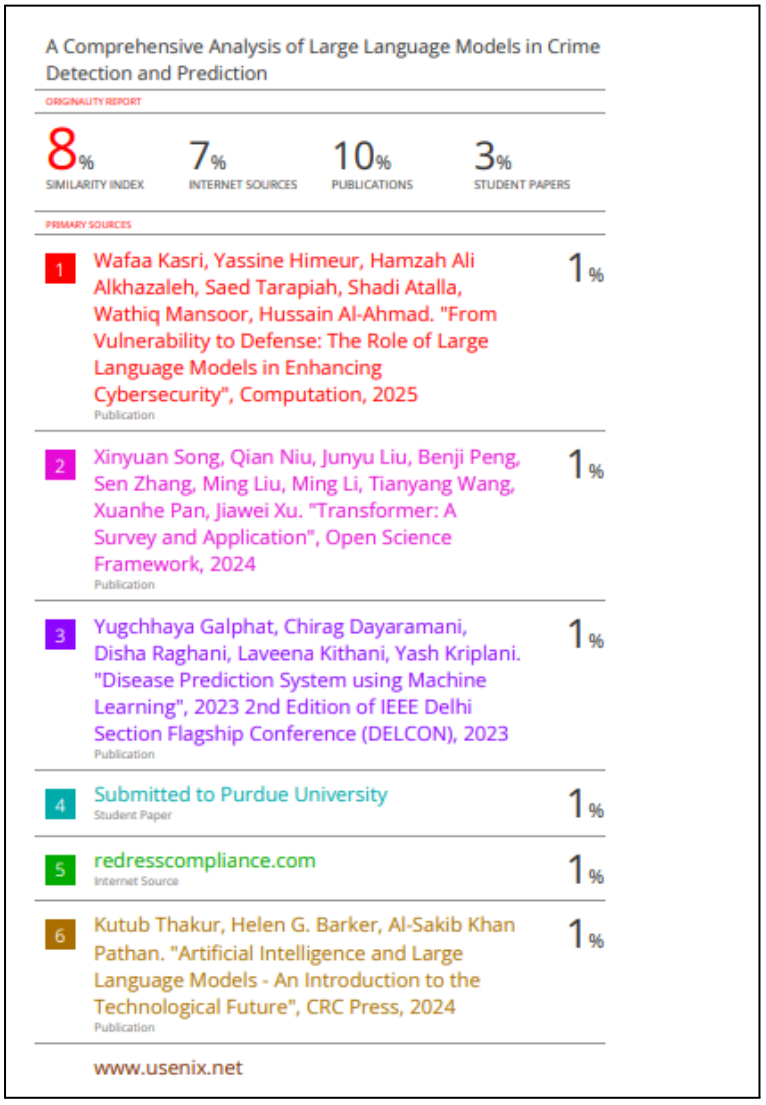
You are requested to give a strict attention to the below mentioned points during preparation of final manuscript to avoid any last minute revert backs from the publisher (Springer).

(a) Please format your paper on the springer template downloaded from our conference website : https://ict4sd.org/ict4sd.php#section08

(b) Authors should not mention designation etc. in the name and affiliation part below the paper title. You are requested to mention only the name, college/organization name, city name and mail ids of all the authors. However, they may mention acknowledgement if necessary (at the end of the paper-prior to references). Kindly mention the full name of your institution and city name.

(c) Authors should not give any photo and biography at the end of the paper.

1.b Plagiarism report of the paper



A Comprehensive Analysis of Large Language Models in Crime Detection and Prediction

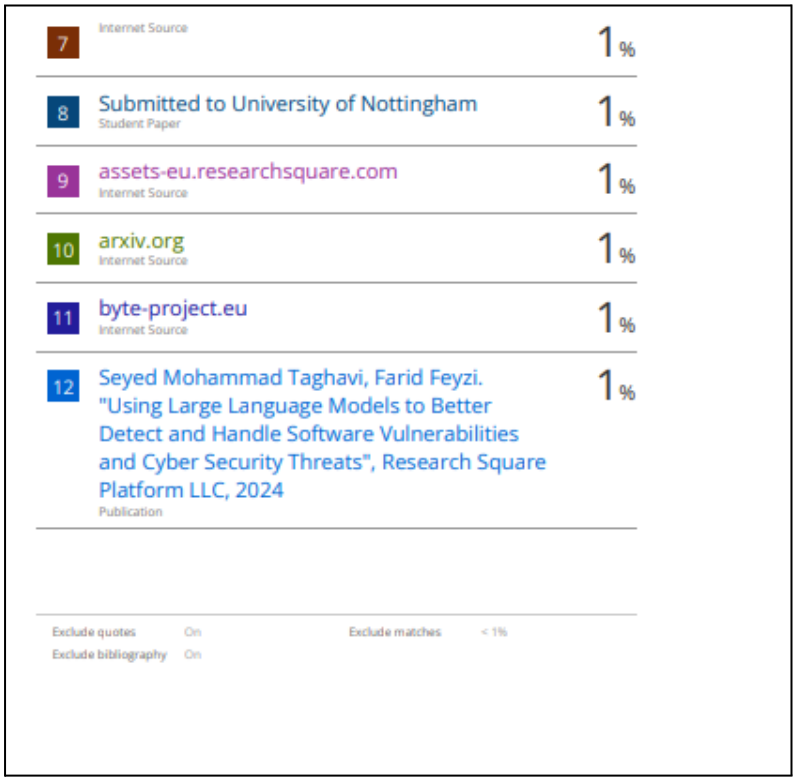
ORIGINALITY REPORT

8%	7%	10%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Wafaa Kasri, Yassine Himeur, Hamzah Ali Alkhazaleh, Saed Tarapiah, Shadi Atalla, Wathiq Mansoor, Hussain Al-Ahmad. "From Vulnerability to Defense: The Role of Large Language Models in Enhancing Cybersecurity", Computation, 2025 Publication	1%
2	Xinyuan Song, Qian Niu, Junyu Liu, Benji Peng, Sen Zhang, Ming Liu, Ming Li, Tianyang Wang, Xuanhe Pan, Jiawei Xu. "Transformer: A Survey and Application", Open Science Framework, 2024 Publication	1%
3	Yugchhaya Galphat, Chirag Dayaramani, Disha Raghani, Laveena Kithani, Yash Kriplani. "Disease Prediction System using Machine Learning", 2023 2nd Edition of IEEE Delhi Section Flagship Conference (DELCON), 2023 Publication	1%
4	Submitted to Purdue University Student Paper	1%
5	redresscompliance.com Internet Source	1%
6	Kutub Thakur, Helen G. Barker, Al-Sakib Khan Pathan. "Artificial Intelligence and Large Language Models - An Introduction to the Technological Future", CRC Press, 2024 Publication	1%

www.usenix.net

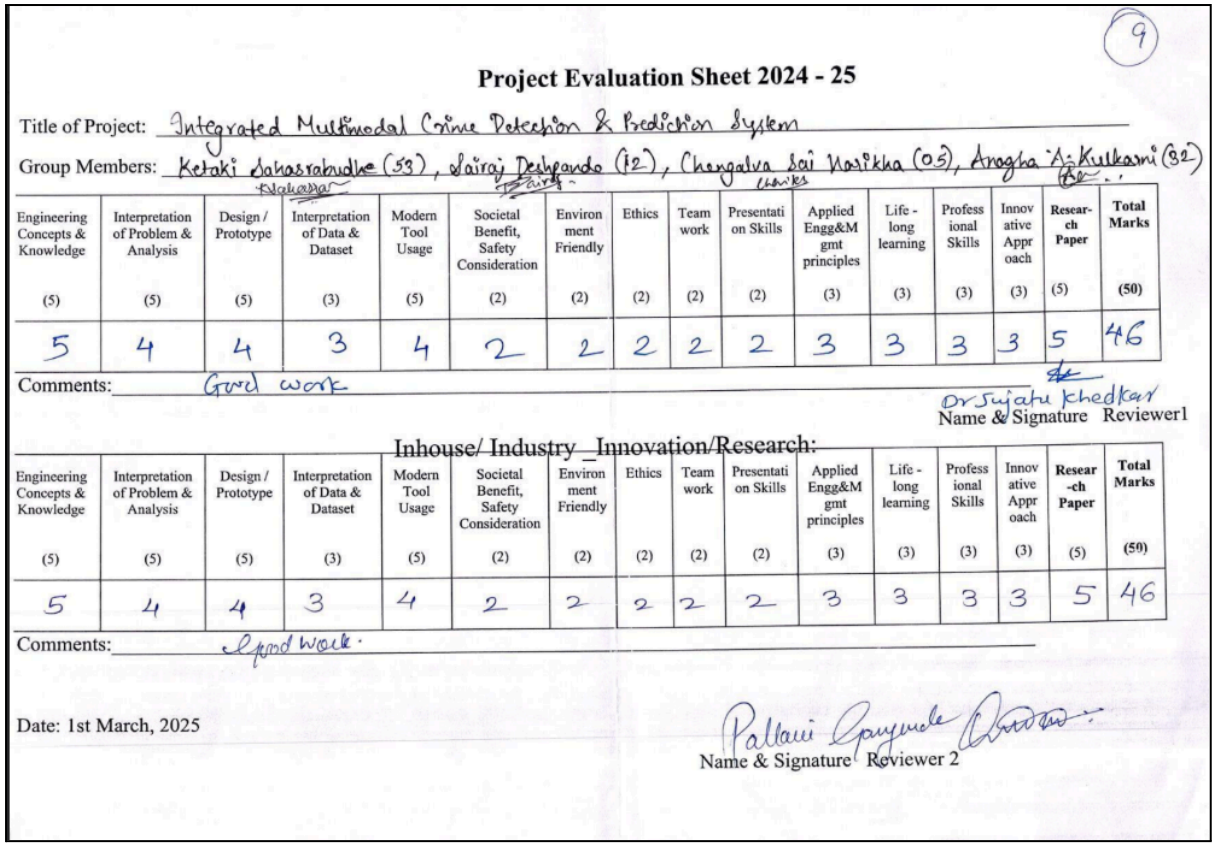


7	Internet Source	1%
8	Submitted to University of Nottingham Student Paper	1%
9	assets-eu.researchsquare.com Internet Source	1%
10	arxiv.org Internet Source	1%
11	byte-project.eu Internet Source	1%
12	Seyed Mohammad Taghavi, Farid Feyzi. "Using Large Language Models to Better Detect and Handle Software Vulnerabilities and Cyber Security Threats", Research Square Platform LLC, 2024 Publication	1%

Exclude quotes On
Exclude bibliography On
Exclude matches < 1%

1.c Project Review sheets

Review 1



Project Evaluation Sheet 2024 - 25

(9)

Title of Project: Integrated Multimodal Crime Detection & Prediction System

Group Members: Ketaki Sahasrabudhe (53), Sairaj Deshpande (12), Chengalva Sai Harikha (05), Anagha A. Kulkarni (32)

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
5	4	4	3	4	2	2	2	2	2	3	3	3	3	5	46

Comments: Good work

Dr Sujata Khedkar
Name & Signature Reviewer1

Inhouse/ Industry _Innovation/Research:

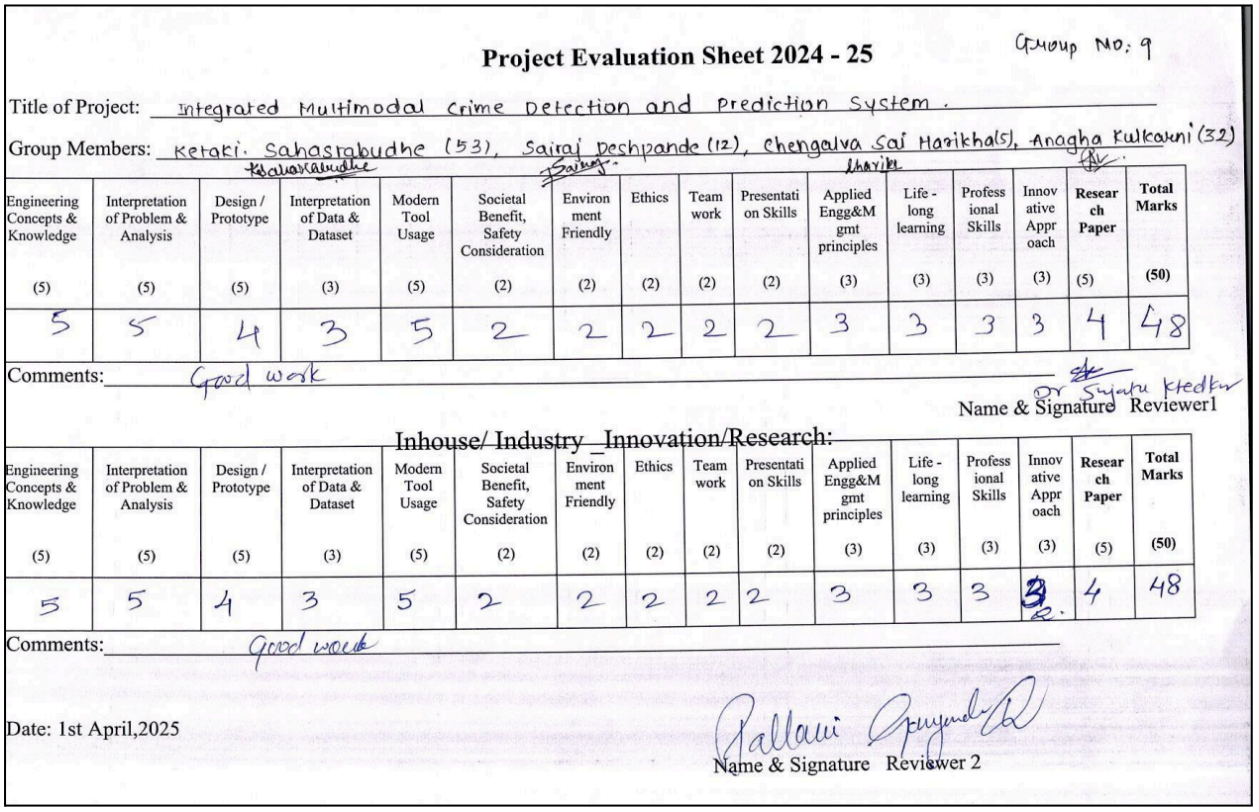
Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
5	4	4	3	4	2	2	2	2	2	3	3	3	3	5	46

Comments: Good work.

Date: 1st March, 2025

Pallavi Apyande
Name & Signature Reviewer 2

Review 2



Project Evaluation Sheet 2024 - 25

Group No: 9

Title of Project: Integrated Multimodal Crime Detection and Prediction System.

Group Members: Ketaki Sahasrabudhe (53), Sairaj Deshpande (12), Chengalva Sai Harikha(5), Anagha Kulkarni (32)

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
5	5	4	3	5	2	2	2	2	2	3	3	3	3	4	48

Comments: Good work

Dr Sujata Khedkar
Name & Signature Reviewer1

Inhouse/ Industry _Innovation/Research:

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
5	5	4	3	5	2	2	2	2	2	3	3	3	3	4	48

Comments: Good work

Date: 1st April,2025

Pallavi Apyande
Name & Signature Reviewer 2