

**VIVEKANAND EDUCATION SOCIETY'S
INSTITUTE OF TECHNOLOGY**

Department of Computer Engineering



Project

Report on

**INTEGRATED MULTIMODAL CRIME DETECTION AND
PREDICTION SYSTEM**

In partial fulfillment of the Fourth Year (Semester–VII), Bachelor of
Engineering

(B.E.) Degree in Computer Engineering at the University of Mumbai

Academic Year 2024-2025

Dr. Mrs. Sujata Khedkar

Submitted by

Chengalva Sai Harikha D12A/ 5

Sairaj Deshpande D12A/12

Anagha Kulkarni D12A/ 32

Ketaki Sahasrabudhe D12A/53

(2024-25)

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 Assistant Professor **Dr. Mrs. Sujata Khedkar** (Project Guide) 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 several times.

Computer Engineering Department**COURSE OUTCOMES FOR B.E PROJECT**

Learners will be to:-

Course Outcome	Description of the Course Outcome
CO 1	Do literature survey/industrial visit and identify the problem of the selected project topic.
CO2	Apply basic engineering fundamental in the domain of practical applications FOR problem identification, formulation and solution
CO 3	Attempt & Design a problem solution in a right approach to complex problems
CO 4	Cultivate the habit of working in a team
CO 5	Correlate the theoretical and experimental/simulations results and draw the proper inferences
CO 6	Demonstrate the knowledge, skills and attitudes of a professional engineer & Prepare report as per the standard guidelines.

Index

Chapter 1: Introduction

- 1.1 Introduction
- 1.2 Motivation
- 1.3 Problem Definition
- 1.4 Existing Systems
- 1.5 Lacuna of the existing systems
- 1.6 Relevance of the Project

Chapter 2: Literature Survey

- A. Brief Overview of Literature Survey
- 2.1 Research Papers Referred
 - a. Abstract of the research paper
 - b. Inference drawn
- 2.2. Inference drawn

Chapter 3: Requirement Gathering for the Proposed System

- 3.1 Introduction to requirement gathering
- 3.2 Functional Requirements
- 3.3 NonFunctional Requirements
- 3.4. Hardware, Software, Technology and tools utilized

Chapter 4: Proposed Design

- 4.1 Block diagram of the system
- 4.2 Modular design of the system

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: Results and Discussion

- 6.1. Performance Evaluation measures
- 6.2. Input Parameters / Features considered
- 6.3. Inference drawn
- 6.4. Screenshots of Code.
- 6.5. Comparison of results with existing systems

Chapter 7: Conclusion

- 7.1 Limitations

7.2 Conclusion

7.3 Future Scope

References

Abstract

In response to the limitations of traditional crime detection methods, this project introduces a crime detection and prediction system that implements the usage of Large Language Models (LLMs) alongside video, audio, and image data. The system integrates advanced machine learning techniques with real-time multimedia analysis and historical crime data to provide a comprehensive solution for crime prevention and law enforcement. By utilising LLMs for analyzing textual data from social media, and other sources, the system enhances contextual understanding and improves the accuracy of crime detection and classification. This approach aims to significantly enhance public safety by uploading the images, text, audio, video to the system. The system is designed to detect the threats that are taking place in the video, audio or the text provided by the user thus, alerting the user about the threat that is taking place and also effectively providing security.

Chapter I

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.

Introduction:

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. Proactive crime prevention can help reduce strain on law enforcement resources and improve response times to high-risk areas.

Chapter II

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.

Literature Survey

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 exclude any posts	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.

		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 multimodal instruction-response pairs . 3) Community Q&A Datasets: Includes data from platforms like Stack Exchange and wikiHow, along with manually created	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.

		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 categories, including a non-crime class.	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 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	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

		Intelligence Graph (FIG) using data from three folders containing three popular Android apps: Phone, Facebook Messenger, and Snapchat		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-responses. • Image instruction dataset: To create an image instruction dataset, we curate around	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

		<p>69K instruction-response pairs by generating them from COCO image captions using GPT-3.5-TURBO as described.</p> <ul style="list-style-type: none"> • Video instruction data: We generate approximately 50K video instruction-response examples by utilizing the video captions from the Charades and AVSD. 		<p>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.</p>
13.	Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding	Video-LLaMA1	Video-LLaMA1	<p>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.</p>
14.	VLM-Eval: A General Evaluation on Video Large Language Models	Video-LLaVA	Video, Text, Images based LLMs.(Video LLama, ImageBind, GPT-4)	<p>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.</p>

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

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.

Requirement Gathering for the Proposed System

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

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.

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.

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.

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.

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

This chapter presents the design framework for the Integrated Multimodal Crime Detection and Prediction System. It includes the system's block diagram, outlining the key components and their interactions, and the modular design, which breaks down the system into manageable and functional modules. These designs serve as the foundation for the system's architecture, ensuring a structured approach to development and implementation.

4 Proposed Design

4.1 Modular Diagram

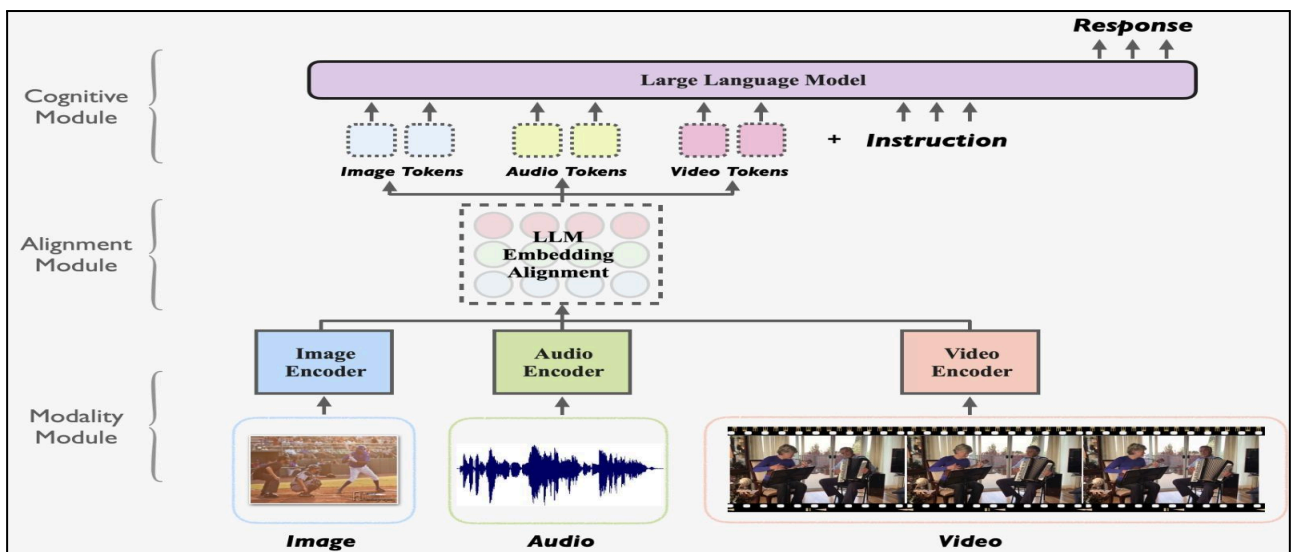


Fig 4.1.1 Modular Diagram 1

4.2 Block Diagram

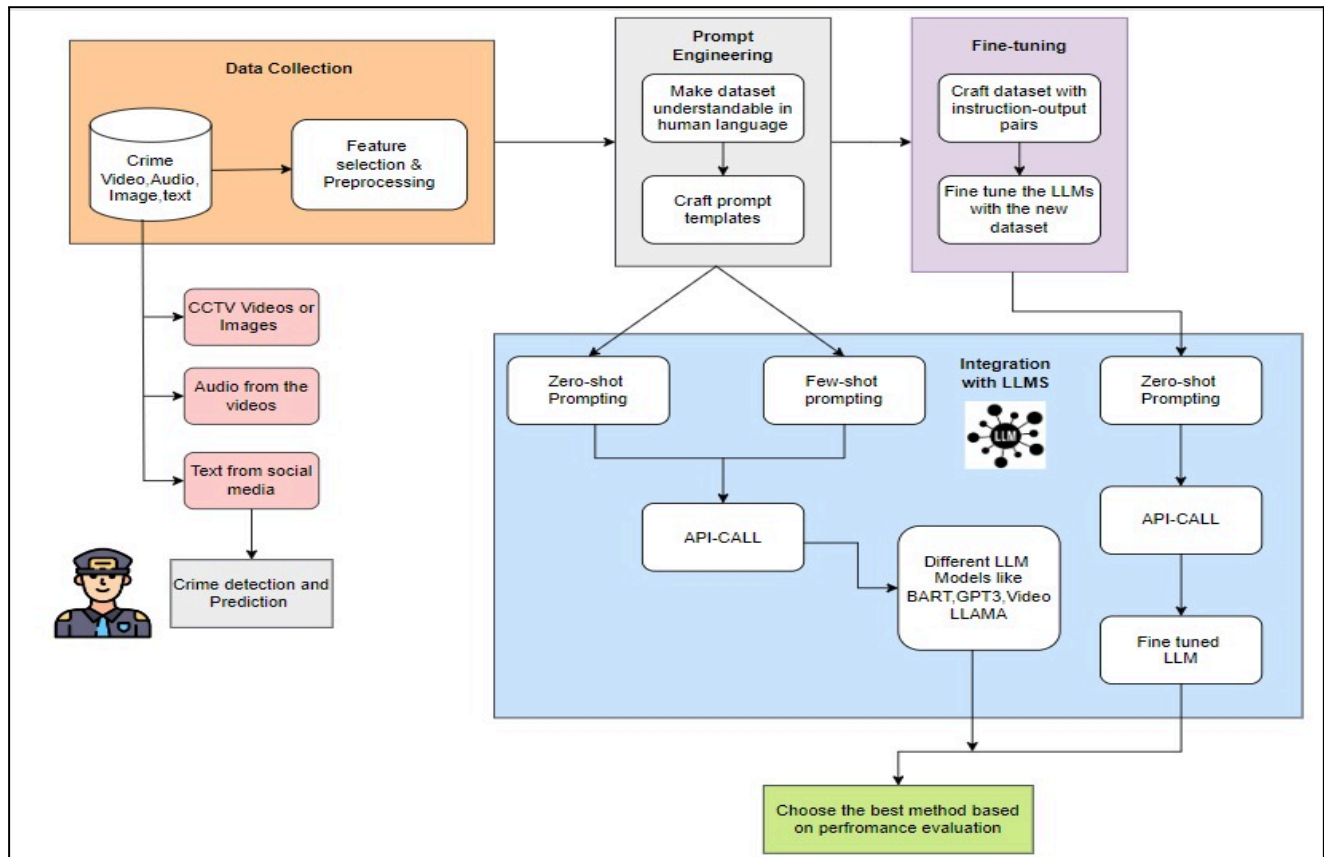


Fig 4.3.1 Block Diagram 1

The above image represents a block diagram for an **Integrated Multimodal Crime Detection and Prediction System** that combines different machine learning and natural language processing techniques to analyze multiple types of data for crime detection and prediction. It has the following components:

1. Data Collection:

- The system collects data from various sources, such as crime videos, audio recordings, images, and text (e.g., from social media).
- These data types are first passed through feature selection and preprocessing to prepare them for further analysis.

2. Prompt Engineering:

- This step involves making the data understandable to large language models (LLMs). The dataset is transformed into a human-readable form through crafting prompt templates.
- These prompts are designed to effectively instruct the model for tasks related to crime detection.

3. Fine-tuning:

- Fine-tuning involves crafting a specialized dataset with instruction-output pairs tailored for crime prediction tasks.
- The LLMs are then fine-tuned with this new dataset to enhance their prediction accuracy for specific crime-related use cases.

4. Integration with LLMs:

- Different types of prompting techniques are integrated into the system:
 - Zero-shot prompting: The LLM is used without any specific task-based training.
 - Few-shot prompting: The LLM is given a small number of examples to guide its predictions.
- API calls are used to connect with various LLM models, such as BART, GPT-3, Video LLAMA, or the fine-tuned LLM.

5. Performance Evaluation:

- The system employs various LLMs and methods (e.g., zero-shot or fine-tuned) and evaluates their performance. The best method is chosen based on this evaluation for the final crime detection and prediction output.

4.3 Data Flow Diagram

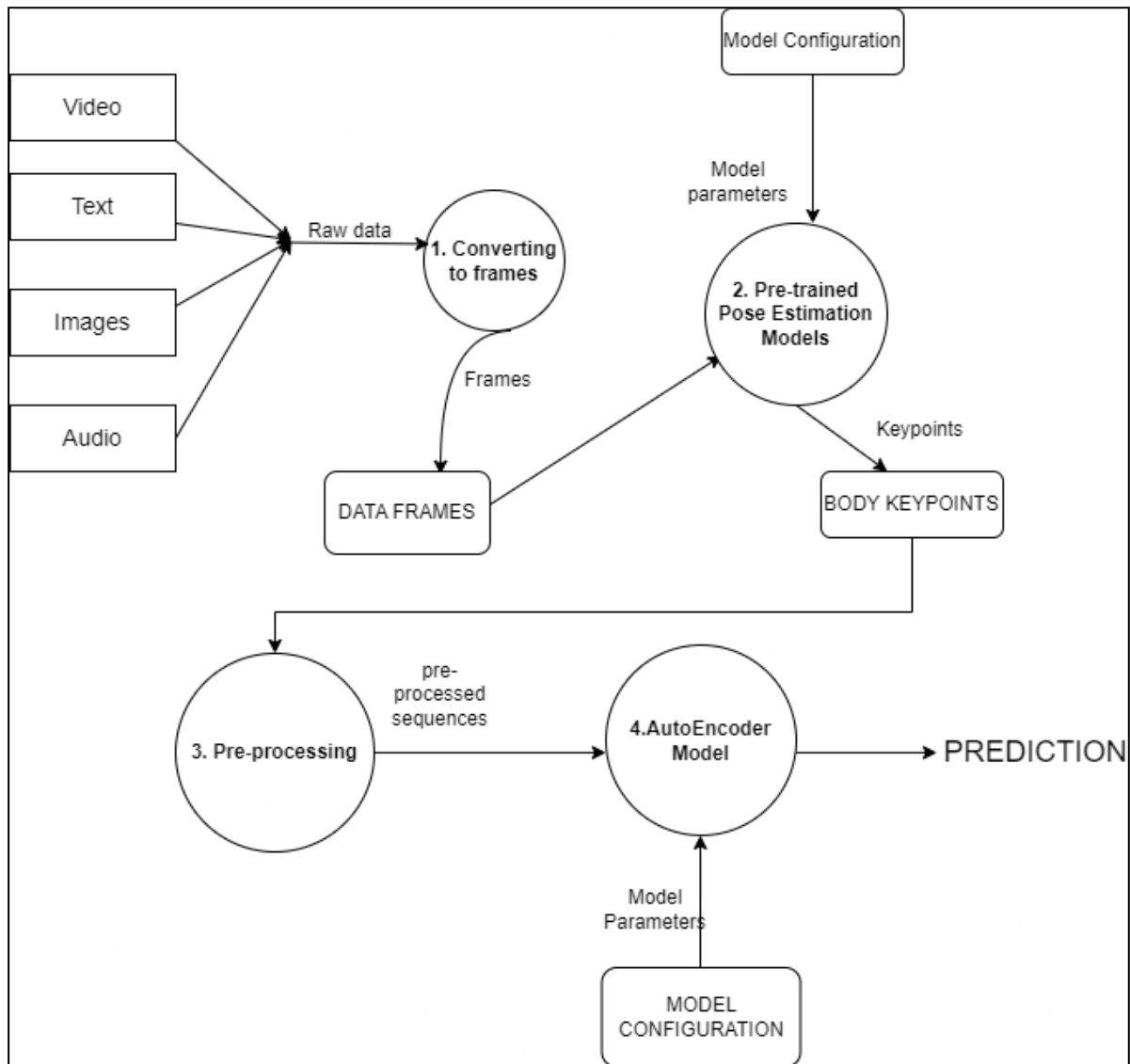


Fig 4.3.1 DFD Diagram 1

The above data flow diagram illustrates the process of crime detection and prevention using multimodal large language models (LLMs). It outlines how different types of data are processed, transformed, and used to make predictions. Here's a breakdown of each part:

1. Raw Data Sources:

The system gathers multimodal data from various sources:

- Video (surveillance footage or crime scene videos),
- Text (social media posts, reports),
- Images (crime scene photos),
- Audio (recordings or voice data).

2. Step 1: Converting to Frames:

All raw data is first converted into a standard format, specifically into frames for

video data or representations for other types of data (images, text, and audio). This conversion allows for uniform processing across different data types.

3. Step 2: Pre-trained Pose Estimation Models:

The frames are then passed through pre-trained pose estimation models, which analyze and extract key features, especially body key points from video or image data. These key points represent specific positions on a person's body that could indicate certain movements or activities relevant to crime detection (e.g., suspicious behavior).

4. Step 3: Pre-processing:

The raw frames and key points extracted from the pose estimation models undergo pre-processing. This step ensures that the data is clean and in the correct format for further analysis. The output is a set of pre-processed sequences ready for model input.

5. Step 4: AutoEncoder Model:

The pre-processed sequences are fed into an AutoEncoder model, which is trained to learn compressed representations of the data. This helps in reducing dimensionality while preserving critical information. The AutoEncoder captures essential patterns in the data that could be indicative of criminal behavior.

6. Model Configuration and Parameters:

Both the pre-trained pose estimation models and the AutoEncoder are fine-tuned using specific model parameters and configurations. These parameters are optimized to ensure the system performs well on crime detection tasks.

7. Prediction:

Finally, the system uses the processed data and learned features to make a prediction. This prediction could relate to detecting a crime in progress or identifying potential future criminal activities, depending on the model's objective.

Chapter V

Implementation of Proposed System

This chapter details the implementation process of the Integrated Multimodal Crime Detection and Prediction System. It describes the methodology used for the system's development, the algorithms and flowcharts that guide each module's functionality, and the datasets sourced and utilized for training and testing. The implementation focuses on transforming the proposed design into a working system that efficiently detects and predicts criminal activities.

5.1 Methodology employed for development

Interaction with Models:

- **Prompting:** The models interacted using prompting techniques via the OpenAI API for GPT models and the Hugging Face repository for BART.
- **Fine-Tuning:** Fine-tuning through instruction tuning was employed to adapt the GPT models to the crime prediction domain.

Prompt Engineering:

- **Importance of Prompts:** The effectiveness of LLMs (Large Language Models) in crime prediction is heavily influenced by the quality of prompts provided.
- **Manual Prompt Engineering:** Manual engineering of prompts was done to design and evaluate the prompts for optimal model performance in crime prediction tasks.

Zero-Shot Learning:

- **Concept:** The models were applied in a zero-shot learning context, where they predict crime categories beyond their explicit training data using semantic relationships and attributes.
- **Prompting in Zero-Shot:** Various templates were used, including prompts with incomplete text or masked slots, allowing the model to predict missing information.
- **Null Prompts:** Simple prompts (null prompts) were utilized, which demonstrated accuracy comparable to manually engineered prompts.

5.2 Datasets' source and utilization

1. Text

- The text data was used initially, consisting of labeled data.
- The csv file had nearly 1000 tuples of labeled data
- Each description of text was associated with a crime type, under the column name "Crime Type"

2. Images

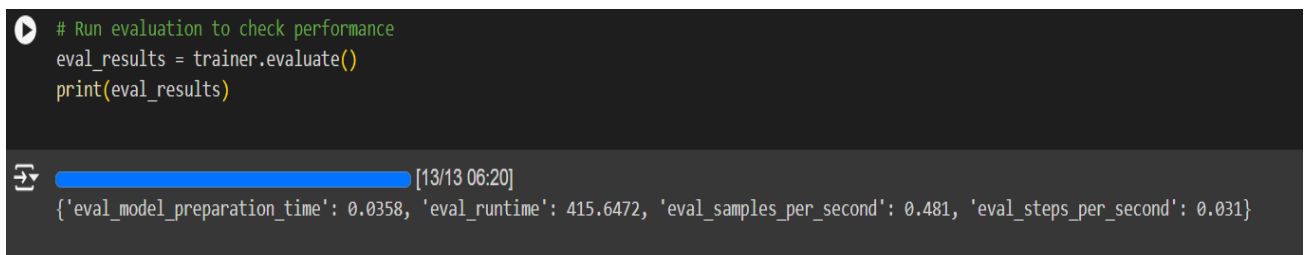
- For images, Smart-City CCTV Violence Detection Dataset (SCVD) was used.
- Current datasets such as the NTU CCTV-Fights dataset, Real-Life Violence Situations dataset (RLVS) and other currently used datasets for violence detection contain videos recorded from phone cameras that could alter the distribution and focus of the CCTV based Violence detection.
- Furthermore, this dataset contains a class for weapons detection in videos, making it the first weapons video dataset as other datasets for weapons detection are based on images of mostly guns and knives.
- This means that the SCVD dataset is tuned to the fact that any handheld object which could be used to harm humans and properties could be regarded as a weapon.

Chapter VI

Results and Discussions

This chapter presents the results obtained from the implementation of the Integrated Multimodal Crime Detection and Prediction System. It includes an evaluation of the system's performance using various metrics, the input parameters or features considered for analysis, and the inferences drawn from the results. Additionally, screenshots of the code are provided for reference, and a comparison is made between the system's results and those of existing crime detection systems to highlight improvements or advancements.

6.1 Performance evaluation measures



```
# Run evaluation to check performance
eval_results = trainer.evaluate()
print(eval_results)
```

[13/13 06:20]

```
{'eval_model_preparation_time': 0.0358, 'eval_runtime': 415.6472, 'eval_samples_per_second': 0.481, 'eval_steps_per_second': 0.031}
```

Fig 6.1.1 Performance evaluation measures

1. Evaluating Model Preparation Time

- Model preparation time refers to the time taken to set up and prepare a machine learning model for training. This can include several steps such as data loading, data preprocessing (cleaning, normalization, feature extraction), and initializing the model architecture and weights.
- In tasks where repeated model training is required, minimizing model preparation time is crucial for improving the overall workflow. A lengthy preparation time can lead to delays in experimentation and model tuning, especially in large-scale systems where datasets are huge, or when hyperparameter optimization techniques (like grid search or random search) are being applied.

2. Evaluating Runtime

- Runtime refers to the amount of time it takes for a machine learning model to complete a specific task, typically during **inference** (making predictions) or **training**. It includes the actual time taken for the computations involved in processing data and updating model parameters during training.
- For real-time systems (e.g., crime detection and prediction), runtime is critical as it dictates the system's responsiveness and efficiency. Reducing runtime without sacrificing accuracy can significantly improve system usability, especially in scenarios where real-time decision-making is essential.

3. Evaluating Samples Per Second

- Samples per second measures how many data samples (e.g., images, text segments, or audio clips) the model processes per second during training or inference. It indicates the model's throughput and can be affected by factors like hardware efficiency (GPU/CPU speed), batch size, and model complexity.
- High samples per second value is desirable as it indicates that the model can process more data in a given time. This is particularly important in scenarios where vast amounts of data are involved, as in multimodal systems like crime detection, where real-time or near-real-time predictions are necessary.

4. Evaluating Steps Per Second

- Steps per second refers to the number of training steps (or iterations) the model completes per second. A **step** in training usually consists of one forward and backward pass over a batch of data, followed by an update of the model parameters. This metric is directly linked to the speed at which the model is trained.
- **Relevance to Performance:** In large-scale systems or models, faster steps per second can reduce overall training time. Optimizing steps per second is essential when experimenting with models, as it directly impacts how quickly the model can converge (i.e., reach the optimal set of parameters) and be deployed for real-world tasks.

6.2 Input Parameters/Features considered

1. Learning Rate

- **Definition:** The learning rate determines the size of the step the model takes when updating its weights during training based on the gradient of the loss function.
- **Effect:**
 - A **high learning rate** can speed up training but might lead to instability or overshooting the optimal solution.
 - A **low learning rate** can make training more stable but slower, with the risk of getting stuck in local minima.
- **Use in LLMs:** It needs to be tuned carefully to balance the training speed and convergence stability. Learning rate schedules (dynamic adjustment) are commonly used to improve performance.

2. Number of Training Steps

- **Definition:** This refers to the total number of updates or iterations the model goes through during the training process.
- **Effect:**
 - **More training steps** allow the model to learn more from the data but can lead to overfitting if the model trains for too long.
 - **Fewer training steps** might lead to underfitting, where the model hasn't learned enough patterns from the data.
- **Use in LLMs:** Adequate training steps ensure that the model learns patterns, syntax, and semantic relationships in the language, improving its ability to generalize.

3. Batch Size

- **Definition:** Batch size refers to the number of samples processed by the model at once before updating the model parameters during training.
- **Effect:**
 - **Larger batch sizes** can speed up training by making better use of hardware (e.g., GPUs), but may require more memory and can lead to less noise in the gradient updates.
 - **Smaller batch sizes** introduce more noise into the gradient estimates, which can help escape local minima but slow down training.

6.3 Inference drawn:

The implementation of an Integrated Multimodal Crime Detection and Prediction System suggests several important inferences. First, the use of multimodal data—such as video surveillance, social media activity, and geographical information—enhances data analysis, leading to improved detection accuracy of crime patterns. Additionally, the application of machine learning algorithms enables predictive capabilities, allowing the system to identify potential crime hotspots and facilitating proactive policing and resource allocation.

Real-time monitoring is another significant advantage, as it enables law enforcement to respond swiftly to incidents as they arise. Moreover, the system promotes interagency collaboration by allowing different law enforcement bodies to share data and insights, resulting in a more coordinated approach to crime prevention.

The system aims to improve public safety by reducing crime rates and increasing the effectiveness of law enforcement operations. However, it also raises ethical considerations regarding privacy and surveillance, necessitating careful attention to data use policies and the importance of maintaining community trust. Furthermore, by optimizing resource utilization through data-driven insights, the system can lead to cost savings and more effective crime-fighting strategies overall.

6.4 Screenshots of Code:

```
!pip install transformers datasets
!pip install torch torchvision torchaudio --extra-index-url https://download.pytorch.org/whl/cu113
```

```
Downloading pyarrow-17.0.0-cp310-cp310-manylinux_2_28_x86_64.whl.metadata (3.3 kB)
Collecting dill<0.3.9,>=0.3.0 (from datasets)
Downloading dill-0.3.8-py3-none-any.whl.metadata (10 kB)
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages (from datasets) (2.1.4)
Collecting xxhash (from datasets)
Downloading xxhash-3.5.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (12 kB)
Collecting multiprocessing (from datasets)
Downloading multiprocessing-0.70.16-py310-none-any.whl.metadata (7.2 kB)
```

```
import pandas as pd
import torch
from transformers import BertTokenizer, BertForSequenceClassification, Trainer, TrainingArguments
from datasets import Dataset
```

```
from google.colab import files

# This will prompt you to upload a file
uploaded = files.upload()
```

No file chosen

Upload widget is only available when the cell has been executed in the current t

Saving crime_dataset_1000.csv to crime_dataset_1000.csv

```
import pandas as pd
```

```
df = pd.read_csv('crime_dataset_1000.csv')
print(df.head())
```

	Crime Description	Crime Type
0	A store was burglarized after hours, and elect...	Burglary
1	A couple was robbed while walking down a deser...	Robbery
2	A fight broke out between two men at a nightcl...	Assault
3	A bike was stolen from the front yard during t...	Theft
4	A woman was forced into a car at knifepoint ne...	Kidnapping

```
from sklearn.model_selection import train_test_split
```

```
train_df, test_df = train_test_split(df, test_size=0.2, random_state=42)
```

```
print(f"Training data size: {len(train_df)}")
print(f"Test data size: {len(test_df)}")
```

```
Training data size: 800
Test data size: 200
```

Implementing the BERT model:

```
tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')

def tokenize_function(examples):
    return tokenizer(examples['Crime Description'], padding='max_length', truncation=True)

tokenized_dataset = dataset.map(tokenize_function, batched=True)
```

/usr/local/lib/python3.10/dist-packages/huggingface_hub/utils/_token.py:89: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (<https://huggingface.co/settings/tokens>)
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models o

warnings.warn(
tokenizer_config.json: 100% ██████████ 48.0/48.0 [00:00<00:00, 2.12kB/s]

vocab.txt: 100% ██████████ 232k/232k [00:00<00:00, 5.08MB/s]

tokenizer.json: 100% ██████████ 466k/466k [00:00<00:00, 6.96MB/s]

config.json: 100% ██████████ 570/570 [00:00<00:00, 23.4kB/s]

/usr/local/lib/python3.10/dist-packages/transformers/tokenization_utils_base.py:1601: FutureWarning:
warnings.warn(
Map: 100% ██████████ 1000/1000 [00:00<00:00, 1109.36 examples/s]

```
# Example input text
input_text = "Bangalore techie steals 50 Laptops from company to cover financial losses"
```

```
inputs = tokenizer(input_text, return_tensors='pt', padding=True, truncation=True)
```

```
print(inputs)
```

```
{'input_ids': tensor([[ 101, 14022, 6627, 2666, 15539, 2753, 12191, 2015, 2013, 2194,
                        2000, 3104, 3361, 6409, 102]]), 'token_type_ids': tensor([[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]]),
```

```
# Predict
predicted_class, probabilities = predict(input_text)

# Print the predicted class name safely
if predicted_class in label_mapping:
    predicted_label = label_mapping[predicted_class]
    print(f"Predicted label: {predicted_label}")
else:
    print(f"Predicted class index {predicted_class} not found in label mapping.")
```

Unique labels in dataset: ['Burglary' 'Robbery' 'Assault' 'Theft' 'Kidnapping' 'Fraud']
Predicted label: Crime Type 3

Chapter VII

Conclusion

This chapter summarizes the overall findings and contributions of the Integrated Multimodal Crime Detection and Prediction System. It discusses the limitations encountered during the development, provides a conclusion on the system's effectiveness in addressing the problem, and outlines potential future enhancements that could improve the system's capabilities or expand its application.

7.1 Limitations:

Computational Resources: Training and fine-tuning BERT can be resource-intensive, requiring substantial computational power and memory, which can be a barrier for many organizations.

Domain Adaptation: While BERT is pre-trained on a large corpus, its performance may drop in specialized domains (like legal texts or specific crime data) unless further fine-tuned on domain-specific data.

Lack of Common Sense: BERT, like many LLMs, lacks true understanding and common sense reasoning, which can lead to errors in inference or predictions.

Bias and Fairness: LLMs can inherit biases present in their training data, leading to biased predictions or outputs that may exacerbate social inequalities.

Interpretability: The "black box" nature of LLMs makes it difficult to interpret their decision-making processes, complicating the evaluation of their outputs.

Resource Intensity: LLMs require significant computational resources for both training and inference, which can limit accessibility and scalability in practical applications.

Dependence on Quality Data: The effectiveness of LLMs heavily relies on the quality and representativeness of the training data; poor data can lead to poor performance.

7.2 Conclusion & Future Scope

LLM-based systems excel in integrating and processing diverse types of data, such as text, audio, images, and video. This multimodal data integration allows these models to form a more comprehensive understanding of complex phenomena like crime. At the core of LLM-based systems is advanced pattern recognition, which uses deep learning techniques to identify trends and anomalies in large datasets.

These systems can detect subtle patterns indicative of criminal behaviour that might be missed by traditional methods. LLM-based systems serve as decision support tools, providing law enforcement agencies with actionable insights. By processing vast amounts of data and delivering concise, relevant information, these systems streamline decision-making processes.

While LLM-based systems offer significant advantages, the theory must also address the ethical implications. Ensuring that these systems operate fairly and without bias is crucial in maintaining public trust.

These theoretical perspectives collectively explain how LLM-based systems enhance crime detection and prediction. By leveraging advanced algorithms, multimodal data, and ethical considerations, these systems support a more proactive, efficient, and just approach to maintaining public safety and preventing crime.

REFERENCES

Simmons, Anj, and Rajesh Vasa. "Garbage in, garbage out: Zero-shot detection of crime using Large Language Models." *arXiv preprint arXiv:2307.06844* (2023). [1]

Sarzaeim, Paria & Mahmoud, Qusay & Azim, Akramul. (2024). A Framework for LLM-Assisted Smart Policing System. IEEE Access. PP. 1-1. 10.1109/ACCESS.2024.3404862. [2]

Kwon, Taeksoo, and Connor Kim. "Efficacy of Utilizing Large Language Models to Detect Public Threat Posted Online." *arXiv preprint arXiv:2401.02974* (2023). [3]

Wu, Fangzhou, et al. "A new era in llm security: Exploring security concerns in real-world llm-based systems." *arXiv preprint arXiv:2402.18649* (2024). [4]

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. [5]

Kumar, Deepak, Yousef AbuHashem, and Zakir Durumeric. "Watch your language: large language models and content moderation." *arXiv preprint arXiv:2309.14517* (2023). [6]

Zhang, Shengyu, et al. "Instruction tuning for large language models: A survey." *arXiv preprint arXiv:2308.10792* (2023). [7]

Henseler, Hans and Harm van Beek. "ChatGPT as a Copilot for Investigating Digital Evidence." *LegalAIIA@ICAIL* (2023). [8]

Achiam, Josh, et al. "Gpt-4 technical report." *arXiv preprint arXiv:2303.08774* (2023). [9]

Ashour Ali, Shahrul Azman Mohd Noah, Lailatul Qadri Zakaria et al. A BERT-Based model: Improving Crime News Documents Classification through Adopting Pre-trained Language Models, 05 March 2023, PREPRINT (Version 1) available at Research Square [<https://doi.org/10.21203/rs.3.rs-2582775/v1>] [10]

"An LLM-driven Approach to Gain Cybercrime Insights with Evidence Networks." [11]

Clairoux-Trepanier, Vanessa, et al. "The Use of Large Language Models (LLM) for Cyber Threat Intelligence (CTI) in Cybercrime Forums." *arXiv preprint arXiv:2408.03354* (2024). [12]

Lyu, Chenyang, et al. "Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration." *arXiv preprint arXiv:2306.09093* (2023). [13]

Zhang, Hang, Xin Li, and Lidong Bing. "Video-llama: An instruction-tuned audio-visual language model for video understanding." *arXiv preprint arXiv:2306.02858* (2023). [14]

Li, Shuailin et al. "VLM-Eval: A General Evaluation on Video Large Language Models." *ArXiv abs/2311.11865* (2023): n. pag. [15]

Munasinghe, Shehan, et al. "PG-Video-LLaVA: Pixel Grounding Large Video-Language Models." *ArXiv* (2023): 2311.13435. [16]

Kong, Zhifeng, et al. "Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities." *arXiv preprint arXiv:2402.01831* (2024). [17]

Wang, Jiaqi, et al. "A comprehensive review of multimodal large language models: Performance and challenges across different tasks." *arXiv preprint arXiv:2408.01319* (2024). [18]

T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 22 199–22 213. [19]