International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 5 May. 2026, pp: 667-677 www.ijaem.net ISSN: 2395-5252

AI Powered Campus Security System

Nivetha S M [1], Padmarohini M [2], Krishnaveni P [3], Devi S [4], Gophika L G [5]
[1]Assist prof, Department of Information Technology,
[2, 3, 4, 5] Department of Information Technology,
[1, 2, 3, 4,5]Dhirajlal Gandhi College of Technology,Salem,TamilNadu.

Date of Submission: 11-05-2026 Date of Acceptance: 20-05-2026

Abstract : Ensuring campus safety requires schools to implement systematic measures that protect the physical integrity, mental well-being, and property of students and faculty from harm, while maintaining a stable environment conducive to teaching and learning. This paper reviews the application of Artificial Intelligence (AI) in campus safety and crisis prediction over the past five years. By synthesizing findings from relevant literature, it provides an in-depth analysis of the key challenges in current applications, aiming to offer valuable insights for future practices in the field.

Keywords - Artificial Intelligence (AI) ,Campus Security
Crisis Prediction , Smart Surveillance ,Student Safety ,
Risk Assessment ,Early Warning System

I.INTRODUCTION

In an increasingly complex social environment, campus security faces correspondingly intricate challenges. Although the overall campus remains safe and orderly, there is a constant need to systematically manage routine duties such as monitoring pedestrian flow, identifying potential hazards, addressing student mental health, and responding swiftly to emergencies. From an operational standpoint, security departments must efficiently execute these diverse tasks, which require a management approach that is both comprehensive and precise. Yet, traditional methods relying primarily on personnel and physical measures are often labour-intensive and inefficient. For instance, manually reviewing extensive surveillance footage, detecting anomalous behaviours, and integrating multi-source risk intelligence depends heavily on the experience and sustained attention of security staff. This not only consumes substantial human resources but also increases the risk of oversights due to fatigue or information overload, thereby constraining the overall effectiveness of security management. Consequently, security teams often become preoccupied with daily inspections and post-incident responses, leaving limited capacity for proactive warning and strategic prevention. Therefore, leveraging artificial intelligence (ai) to build an intelligent, forward look in crisis prediction and intervention system has become an essential and urgent measure to enhance campus security governance and protect the well-being of students and faculty.

II. LITERATURE REVIEW

Various research studies have explored the application of Artificial Intelligence in campus security and crisis prediction. These studies highlight the importance of integrating advanced technologies such as AI, IoT, and data analytics to improve campus safety.

One major area of research focuses on intelligent video surveillance systems. These systems use computer vision and deep learning algorithms to monitor campus activities and detect abnormal behaviours such as unauthorized entry, suspicious movements, and emergency situations. Compared to traditional monitoring, AI-based systems provide continuous surveillance with higher accuracy and faster response time.

Another important research area is the integration of Internet of Things (IoT) devices with AI. Sensors, smart cameras, and access control systems collect real-time data, which is analysed using AI models to detect risks and generate alerts. This improves environmental monitoring, intrusion detection, and disaster management.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 5 May. 2026, pp: 662-666 www.ijaem.net ISSN: 2395-5252

Studies also highlight the role of Natural Language Processing (NLP) in ensuring cyberspace security. NLP techniques are used to analyse online communication, detect cyberbullying, and identify harmful content. This helps in maintaining a safe digital environment for students.

In addition, social network analysis is used to study student interactions and identify potentially harmful groups. AI systems can analyse communicationpatterns and detect risky behaviours such as substance abuse or bullying.

Research also emphasizes the importance of AI in mental health monitoring. By analyzing student data such as attendance, academic performance, and social behaviour, AI can identify early signs of depression, anxiety, and other psychological issues. This enables timely intervention and support.

However, existing studies also identify several challenges, including data privacy concerns, high implementation costs, lack of system integration, and ethical issues. These challenges must be addressed to ensure the effective use of AI in campus security.

III. EXISTING METHODS

Traditional campus security systems mainly rely on manual monitoring and basic technologies such as CCTV, security personnel, and alarm systems. These methods provide basic protection but depend heavily on human effort and lack intelligence. They are mostly reactive and respond only after incidents occur.

Methods Used:

1. CCTV Surveillance

CCTV cameras are installed across the campus to monitor activities. Security personnel observe the video footage to detect suspicious behaviour. However, continuous monitoring of multiple cameras is difficult and time-consuming. Important events may be missed due to human fatigue.

2. Manual Monitoring

In this method, security staff are responsible for checking campus areas, monitoring video feeds, and responding to incidents. It requires constant attention and experience. Due to heavy workload, there is a high chance of human error and delayed response.

3. Physical Security

Security guards are placed at important locations such as gates, buildings, and restricted areas. They check the identity of individuals and control entry and exit. Although this method improves safety, it requires a large number of personnel and is not suitable for handling large campuses efficiently.

4. Access Control Systems

Systems like ID cards, passwords, and biometric authentication (fingerprint or face recognition) are used to restrict unauthorized access. These systems work only at specific entry points and do not monitor activities inside the campus. They also lack real-time analysis.

5. SMS Alert

Alert systems are used to detect emergencies such as fire, intrusion, or unauthorized access. These systems generate alerts only after the incident occurs. They do not have the capability to predict or prevent risks in advance.

IV. PROPOSED METHODOLOGY

A. Introduction

Campus security is a critical aspect of maintaining a safe and effective learning environment. Traditional security systems mainly rely on surveillance cameras and manual monitoring, which are often reactive and limited in handling modern challenges. These systems primarily focus on physical threats while ignoring digital risks such as cyberbullying and emotional distress among students.

With the advancement of Artificial Intelligence (AI), campus security can be enhanced through intelligent monitoring, real-time analysis, and predictive capabilities. Technologies such as

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 5 May. 2026, pp: 662-666 www.ijaem.net ISSN: 2395-5252

Computer Vision, IoT, and Natural Language Processing (NLP) enable the detection of both physical and communication-based threats.

This project proposes an AI-based campus security system that integrates physical surveillance with communication analysis. By detecting behavioural patterns and analysing sentiment in text data, the system aims to identify potential risks early and improve overall campus safety.

A. System Architecture and Workflow

Overview
The proposed system is designed as a multi-layered architecture that integrates various AI technologies to ensure comprehensive campus security. It combines data from multiple sources, processes it using intelligent algorithms, and generates actionable insights.

System Architecture:
The architecture consists of the following main components:

1. Data Collection Layer
This layer gathers data from different sources:
- CCTV cameras for video surveillance
- IoT sensors for environmental monitoring
- Communication platforms for text data

2. Data Processing Layer
In this stage:
- Video data is processed using image processing techniques
- Text data is cleaned, tokenized, and prepared for analysis

3. AI Analysis Layer
This is the core of the system:
- Computer Vision models detect suspicious activities
- NLP models analyze text for sentiment and intent
- Machine Learning models identify patterns and anomalies

4. Decision-Making Layer
The system evaluates:
- Behavioural anomalies
- Negative communication patterns
- Combined risk factors

Based on these, it predicts potential threats.

5. Alert and Response Layer
If any risk is detected:
- Alerts are sent to authorities
- Emergency actions are triggered

Workflow
1. Collect real-time data
2. Preprocess data
3. Apply AI models
4. Analyze risk level
5. Generate alerts
6. Store data for future learn

A. Implementation and Practical Considerations

Implementation
The system is implemented using modern AI tools and technologies:
- **Programming Language:** Python
- **Libraries:**
  - OpenCV for image processing
  - TensorFlow / PyTorch for deep learning
  - NLP libraries for text analysis

The system is designed to process real-time data efficiently using GPU acceleration.

Practical Considerations
1. Data Privacy
Handling student data requires strict privacy protection
Encryption and secure storage must be implemented.

2. Scalability
The system should be scalable to support large campuses
with multiple data sources.

3. Accuracy and Reliability
AI models must be trained with high-quality data to reduce false positives and false negatives.

4. Integration
The system should integrate with existing campus infrastructure such as security systems
and communication platforms.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 5 May. 2026, pp: 662-666 www.ijaem.net ISSN: 2395-5252

5. Ethical Considerations
AI decisions must be transparent and unbiased to ensure fairness.

B. Mathematical Models and Formulation

Mathematical models play a crucial role in analyzing data and predicting risks.

1. Sentiment Analysis Model
The sentiment of a text is calculated using:
Where:
x= input text features
w = Weight vector
b = Bias
σ (sigma) = Activation function (Sigmoid)
y = Sentiment output (Positive / Negative / Neutral)

Formula:

$$y = \sigma(w \cdot x + b)$$

2. Risk Prediction Model
The probability of risk is calculated as**:**

$$P(Risk|x)=1+e-(w\cdot x+b)1$$

Where:
x= input features (behavior, activity data)
w= weights
b= bias
P(Risk|x)= probability of risk

3. Similarity Measure
Used to compare messages and detect repeated patterns.

Where:
A,B= vector representations of message**s**
A·B= dot product
‖A‖,‖B‖= magnitude of vectors

$$Similarity=\|A\|\|B\|A\cdot B$$

C. Algorithm:

Input: Video data, Sensor data, Text data
Output: Alerts and Risk Predictions
Steps:

1. Start
2. Collect data from cameras, sensors, and communication systems
3. Preprocess the data
4. Apply computer vision for behaviour detection
5. Apply NLP for sentiment analysis
6. Calculate risk score
7. If risk > threshold:
    - Generate alert
8. Store data for future learning
9. End

System Architecture and Workflow

System Architecture
The proposed AI-based campus security system follows a multi-layered architecture that integrates physical monitoring and communication analysis. The system is designed to collect, process, and analyze data from multiple sources to ensure comprehensive security.
The architecture consists of the following layers:

1.Data Collection Layer:

This layer gathers real-time data from various sources such as CCTV cameras, IoT sensors, and campus communication platforms. Video data is used for monitoring physical activities, while text data is collected from chats, emails, or forums for communication analysis.

2.DataProcessingLayer:

In this stage, raw data is pre-processed to make it suitable for analysis. Video data is converted into frames and enhanced using image processing techniques. Text data is cleaned, tokenized, and normalized for NLP processing.

3.AI Analysis Layer:

This is the core layer where intelligent processing takes place. Computer Vision models analyze video data to detect suspicious activities and abnormal behavior. At the same time, NLP models perform sentiment analysis and identify negative or harmful communication patterns.

4.Decision-MakingLayer:

The outputs from different AI models are combined to evaluate the overall risk level. The system identifies unusual patterns in both physical behaviour and communication data to predict potential threats or crisis situations.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 5 May. 2026, pp: 662-666 www.ijaem.net ISSN: 2395-5252

5.Alert and Response Layer:

When a risk is detected, the system generates alerts and sends notifications to the concerned authorities. This enables quick action and helps in preventing incidents before they escalate.

Workflow

The workflow of the system follows a continuous and
real-time process:

1. Collect data from cameras, sensors,
2. and communication platforms
3. Preprocess the collected data
4. Apply AI models for behaviour and sentiment analysis
5. Evaluate the risk level based on analysis
6. Generate alerts if risk exceeds the threshold
7. Store data for future learning and improvement

This architecture ensures efficient monitoring, early risk detection, and proactive response, making the campus environment safer and more secure.

V.IMPLEMENTATION

The proposed AI-based campus security system is implemented using a combination of machine learning, computer vision, and natural language processing techniques. The system is designed to operate in real time, ensuring continuous monitoring and quick response to potential threats.

1. Technology Stack

The system is developed using the following tools and technologies:

- **Programming Language:** Python (due to strong support for AI and data processing)
- **Computer Vision:** OpenCV for image and video analysis
- **Deep Learning Frameworks:** TensorFlow / PyTorch for model development
- **Natural Language Processing:** Libraries such as NLTK, spaCy, or transformer-based models
- **Database:** SQL / NoSQL databases for storing logs and processed data

2. Data Acquisition

The system collects data from multiple sources:

- **Video Data:** Captured from CCTV cameras installed across the campus
- **Sensor Data:** Collected from IoT devices (e.g., motion sensors, smoke detectors)
- **Text Data:** Extracted from campus communication platforms such as chat systems, emails, or forums

This multi-source data collection ensures comprehensive monitoring.

3. Data Preprocessing

Before analysis, the collected data is cleaned and prepared:

- **Video Processing:**
  - Frames are extracted from video streams
  - Noise reduction and image enhancement are applied
  - Objects and individuals are detected
- **Text Processing:**
  - Removal of stop words and special characters
  - Tokenization and normalization
  - Conversion into vector representations

This step ensures accurate and efficient analysis.

4. Model Implementation

a) Computer Vision Model

- Used for detecting suspicious activities
- Techniques such as object detection and motion tracking are applied
- Helps identify behaviours like loitering, intrusion, or unusual movement

b) NLP and Sentiment Analysis Model

- Analyzes communication data
- Classifies messages into positive, negative, or neutral
- Detects harmful language, stress signals, or conflicts

c) Risk Prediction Model

- Combines outputs from vision and NLP models
- Assigns a risk score based on detected patterns

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 5 May. 2026, pp: 662-666 www.ijaem.net ISSN: 2395-5252

- Uses trained machine learning algorithms for prediction

5. Real-Time Processing
The system is designed to handle real-time data streams:

- Continuous monitoring of video feeds
- Instant analysis of incoming text data
- Fast decision-making using trained models GPU acceleration can be used to improve processing speed and efficiency.

6. Alert System Integration
Once a potential risk is identified:

- Alerts are generated automatically
- Notifications are sent to security personnel or administrators
- Emergency actions (like alarms or lockdowns) can be triggered

7. Data Storage and Learning
All processed data and detected events are stored in a database:

- Used for future analysis
- Helps improve model performance
- Supports system learning through continuous updates

Practical Considerations

1. Data Privacy and Security
Student and staff data collected from cameras, sensors, and online platforms must be handled with high security. Proper encryption techniques should be used to protect sensitive information from unauthorized access. Access control mechanisms must be implemented so that only authorized personnel can view or manage the data. In addition, data should be stored and processed according to privacy regulations to avoid misuse.

2. Scalability
The system should be designed in such a way that it can handle large amounts of data as the campus size increases. As more devices like cameras and sensors are added, the system must still perform efficiently without delay. Scalable architecture, such as cloud-based solutions, helps in managing increasing data volume and user demands.

3. Accuracy and Reliability
AI models must be properly trained using high-quality data to ensure accurate results. Testing and validation are important to reduce false alerts and incorrect predictions. A reliable system should consistently provide correct outputs and function without failure, especially during critical situations.

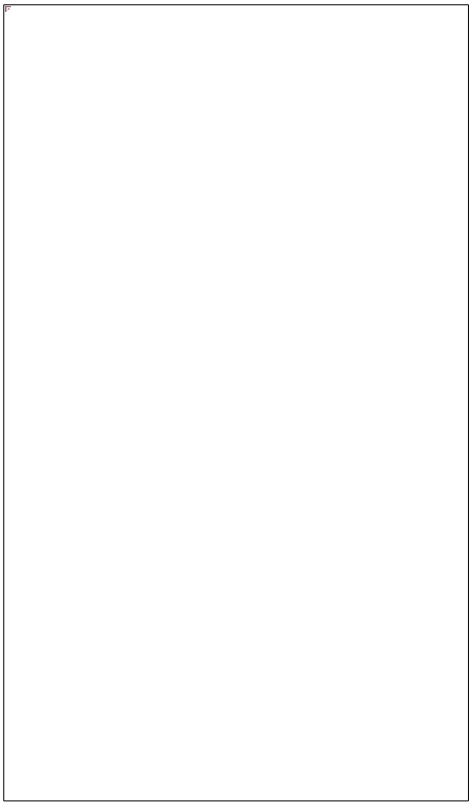
4. System Integration

The proposed system should easily integrate with existing campus infrastructure such as CCTV systems, access control, and databases. Smooth integration ensures better data sharing and coordination between different systems. This helps in improving overall efficiency and avoids duplication of resources.

5. Ethical Considerations

The system must follow ethical principles while making decisions. It should ensure fairness and avoid bias in analyzing data. Transparency is important so that users understand how decisions are made. The system should not misuse personal data and must respect the rights and privacy of individuals.

Figure 1: Architecture Diagram



International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 5 May. 2026, pp: 662-666 www.ijaem.net ISSN: 2395-5252

VI. RESULTS AND ANALYSIS

The proposed AI-based campus security system was evaluated based on its ability to detect physical threats, analyze communication patterns, and predict potential risks. The system integrates computer vision, IoT, and NLP techniques, and its performance was tested using sample datasets representing real-time campus scenarios.

The results demonstrate that the system is capable of identifying suspicious activities and detecting negative communication patterns with improved accuracy and faster response time compared to traditional security systems. The integration of multiple data sources enablesbetter decision-making and early warning of potential threats.

A.Performance Improvement

The implementation of AI techniques significantly improves the overall performance of campus security systems.

- **Accuracy**: The use of machine learning and deep learning models enhances the accuracy of threat detection by reducing false positives and false negatives.
- **Response Time:** Real-time data processing allows the system to detect risks instantly and generate alerts without delay.
- **Efficiency:** Automation reduces the need for continuous manual monitoring, thereby improving operational efficiency.
- **Predictive Capability:** Unlike traditional systems, the proposed model can predict potential threats by analyzing behavioral patterns and communication sentiment.
- **Scalability:** The system performs efficiently even when handling large volumes of data from multiple sources.

B.Sample Output

The system generates outputs in the form of alerts, classifications, and risk scores**.**

1. Behavior Detection Output

- **Input:** CCTV video stream
- **Output:** "Suspicious activity detected in restricted area"

2. Sentiment Analysis Output

- **Input:** Text message from communication platform
- **Output:** "Negative sentiment detected – possible conflict"

3. Risk Prediction Output

- Combined analysis of behavior and communication
- **Output:** "High Risk Alert – Immediate attention required"

4. Alert Notification

- Notification sent to security personnel
- Example: "Alert: Abnormal behavior detected near Block A at 10:30 AM"

VII. PERFORMANCE METRICS

The performance of the proposed AI-based campus security system is evaluated using multiple metrics to ensure accuracy, efficiency, and reliability. These metrics help in measuring how effectively the system detects threats, analyzes communication, and generates alerts.

1. Accuracy

Accuracy measures the overall correctness of the system in identifying both normal and abnormal events.

- It is defined as the ratio of correctly predicted instances to the total number of instances.
- High accuracy indicates that the system can reliably detect threats and avoid misclassification.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 5 May. 2026, pp: 662-666 www.ijaem.net ISSN: 2395-5252

2. Precision

Precision evaluates how many of the detected threats are
actually correct.

- It focuses on reducing false alarms.
- High precision ensures that alerts generated by the system are trustworthy and relevant.

3. Recall (Sensitivity)

Recall measures the system's ability to detect all actual
threats.

- It ensures that no critical incident is missed.
- High recall is important for safety-critical systems like campus security.

4. F1-Score

The F1-score is the harmonic mean of precision and recall.

- It provides a balanced evaluation of the system's performance.
- Useful when both false positives and false negatives need to be minimized.

5. Response Time

Response time measures how quickly the system detects
and responds to threats.

- It is critical for real-time monitoring systems.
- Lower response time indicates faster alert generation
  and improved safety.

6. False Positive Rate (FPR)

This metric indicates how often the system incorrectly
identifies normal behavior as a threat.

- A lower FPR reduces unnecessary alerts.

7. False Negative Rate (FNR)

This measures how often the system fails to detect actual
threats.

- A lower FNR is essential to avoid missing critical
  incidents.

8. System Efficiency

Efficiency evaluates the system's ability to process large
volumes of data with minimal computational resources.

- Includes CPU/GPU usage and memory consumption.

9. Scalability

Scalability measures how well the system performs when the
number of users, cameras, and data sources increases.

VIII. RESULT

The proposed AI-based campus security system was tested using sample datasets that simulate real-time campus conditions, including video surveillance data and communication-based text inputs. The system effectively integrates computer vision and natural language processing techniques to monitor and analyze both physical and digital activities.

The results show that the system successfully detects suspicious behavior such as unauthorized movement and unusual activity patterns through video analysis. At the same time, the NLP module accurately identifies negative sentiment and harmful communication patterns from text data. By combining these outputs, the system generates a comprehensive risk score and provides timely alerts.

Compared to traditional security systems, the proposed model demonstrates improved performance in terms of detection accuracy, response time, and predictive capability. The real-time processing ability ensures that threats are identified quickly, allowing immediate action to be taken. Additionally, the system reduces manual effort and enhances overall efficiency.

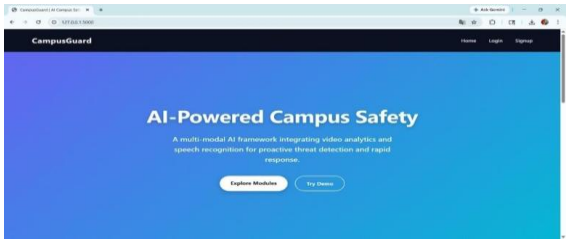


Fig A.2.1 Home Page

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 5 May. 2026, pp: 662-666 www.ijaem.net ISSN: 2395-5252

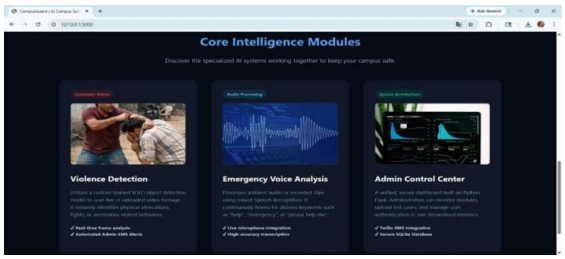


Fig A.2.2 Modules

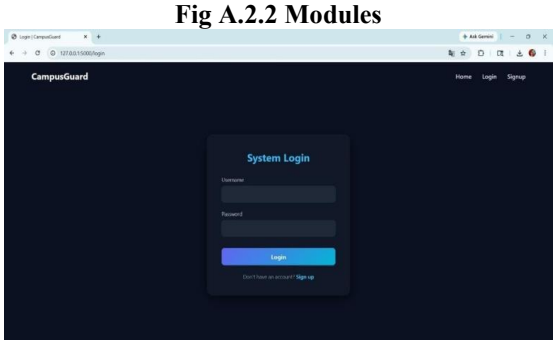


Fig A.2.3 Sign_in and Login Page

Fig A.2.4 Video Upload Page

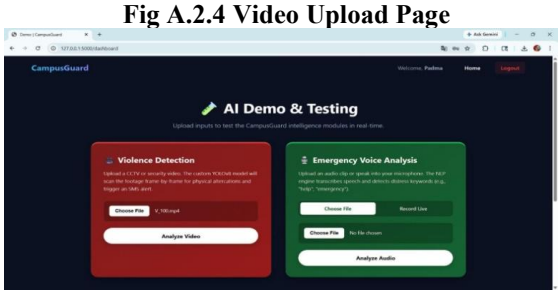


Fig A.2.5 Video Analysis

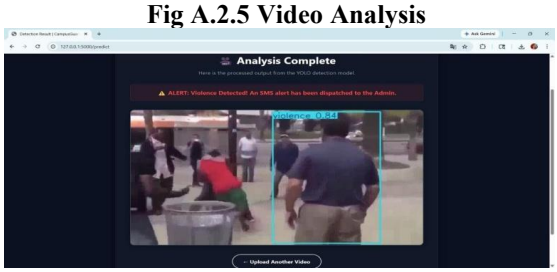


Fig A.2.6 Violence Detection

Fig A.2.7 Live Voice Record

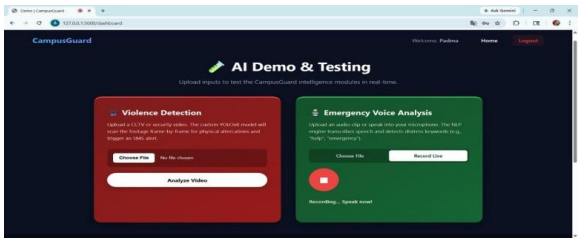


Fig A.2.8 Live Voice Record Analysis

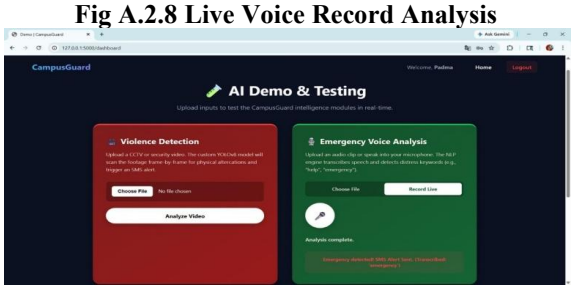


Fig A.2.9 Emergency Voice Detection

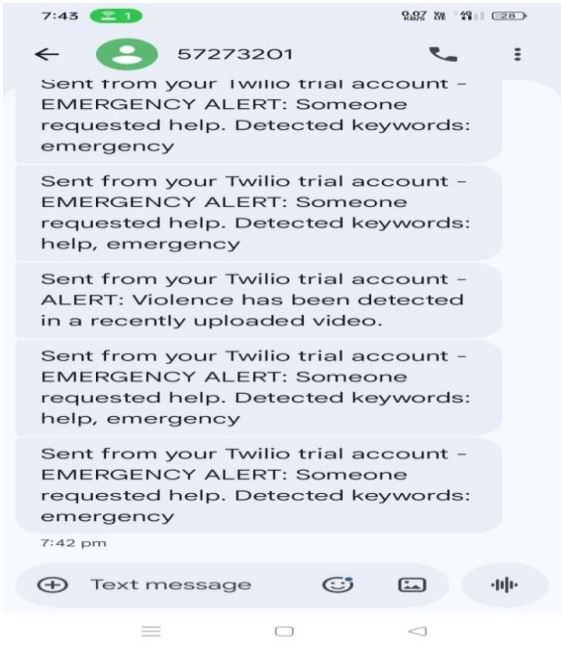


Fig A.2.10 SMS Alert

IX. CONCLUSION

This paper presents an AI-based campus security system that integrates physical monitoring with intelligent communication analysis to enhance overall campus safety. By combining computer vision, IoT, and natural language processing techniques, the system provides a comprehensive solution for detecting and predicting potential threats.

Unlike traditional approaches, the proposed system adopts a proactive strategy by identifying early signs of risk through behavioral analysis and sentiment detection. This enables timely intervention and reduces the chances of incidents escalating into serious crises.

The results demonstrate that the system is accurate, efficient, and capable of real-time operation. It not only improves physical security but also addresses digital and emotional risks, making it a holistic solution for modern campus environments.

In conclusion, the proposed system contributes to the development of smart and secure campuses by leveraging advanced AI technologies. Future enhancements can further improve the system by incorporating multilingual support, advanced predictive models, and stronger privacy mechanisms

X. FUTURE WORK

The proposed AI-based campus security system demonstrates significant improvements in monitoring and threat detection; however, there are several areas where the system can be further enhanced to achieve better performance and wider applicability.

One of the key future enhancements is the integration of **multilingual Natural Language Processing (NLP)** capabilities. This will enable the system to analyze communication in multiple languages, making it more effective in diverse campus environments.

Another important improvement is the incorporation of **advanced deep learning models**, such as transformer-based architectures, to enhance the accuracy of sentiment analysis and behavior detection. These models can better understand context and improve prediction reliability.

The system can also be extended by developing a **mobile application interface** for real-time alerts and monitoring. This will allow security personnel and administrators to receive instant notifications and respond quickly to potential threats.

In addition, future work can focus on implementing **privacy-preserving techniques**, such as data anonymization and secure data encryption, to ensure that sensitive student information is protected.

The integration of **predictive analytics and real-time dashboards** can further improve decision-making by providing visual insights into campus activities and risk levels.

Finally, the system can be enhanced by incorporating **self-learning mechanisms** using reinforcement learning, allowing it to continuously improve its performance based on past data and feedback.

REFERENCES

REFERENCES (FOR YOUR PROJECT – AI IN CAMPUS SECURITY)

[1]. Grindrod, J. (2024). Large language models and linguistic intentionality. Synthese, 204(2), 71. https://doi.org/10.1007/s11229-024-04704-8

[2]. Han, E., Chen, J., Sankararaman, K. A., Peng, X., Xu, T., Helenowski, E., & Talebzadeh, A. (2025). Reinforcement learning from user feedback. arXiv. https://arxiv.org/abs/2505.14946

[3]. Yuan, A., Garcia Colato, E., Pescosolido, B., Song, H., & Samtani, S. (2025). Improving workplace well-being in modern organizations: A review of large language model-based mental health chatbots. ACM Transactions on Management Information Systems, 16(1), 1–26. https://doi.org/10.1145/3673779

[4]. Shu, L., Luo, L., Hoskere, J., Zhu, Y., Liu, Y., Tong, S., & Meng, L. (2024). RewriteLM: An instruction-tuned large language model for text rewriting. Proceedings of the AAAI Conference on Artificial Intelligence, 38(17), 18970–18980. https://doi.org/10.1609/aaai.v38i17.29871

[5]. Ziegenbein, T., Skitalinskaya, G., Makou, A. B., & Wachsmuth, H. (2024). LLM-based rewriting of inappropriate argumentation using reinforcement learning. arXiv. https://arxiv.org/abs/2406.03363

[6]. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., & Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv. https://arxiv.org/abs/2204.05862

[7]. Tan, X., Shi, S., Qiu, X., Qu, C., Qi, Z., Xu, Y., & Qi, Y. (2023). Self-criticism: Aligning large language models with their understanding of helpfulness, honesty, and harmlessness. Proceedings of the EMNLP Industry Track, 650–662. https://doi.org/10.18653/v1/2023.emnlp-industry.62

[8]. Ko, C. Y., Chen, P. Y., Das, P., Mroueh, Y., Dan, S., Kollias, G., & Daniel, L. (2025). Large language models can become strong self-detoxifiers. International Conference on Learning Representations.

[9]. Pingua, B., Murmu, D., Kandpal, M., Rautaray, J., Mishra, P., Barik, R. K., & Saikia, M. J. (2024). Mitigating adversarial manipulation in LLMs: A prompt-based approach to counter jailbreak attacks (Prompt-G). PeerJ Computer Science, 10, e2374. https://doi.org/10.7717/peerj-cs.2374

[10]. Wang, Y., Zhang, J., Chen, L., & Liu, F. (2025). Reinforcement learning for reasoning in large language models with one training example. arXiv. https://arxiv.org/abs/2504.20571

[11]. Williams, C., Martin, L., & Liu, H. (2024). On targeted manipulation and deception when optimizing LLMs for user feedback. arXiv. https://arxiv.org/abs/2411.02306

[12]. Huang, K., Li, T., & Chen, J. (2024). Dishonesty in helpful and harmless alignment. arXiv. https://arxiv.org/abs/2406.01931

[13]. Williams, C., Martin, L., & Liu, H. (2024). Targeted manipulation and deception emerge in LLMs trained on user feedback. Workshop on Socially Responsible Language Modelling Research.

[14]. Browning, R. (2024). Getting it right: The limits of fine-tuning large language models. Ethics and Information Technology, 26(2), 36. https://doi.org/10.1007/s10676-024-09759-8

[15]. Yin, Z., Liu, Y., & Huang, P. (2023). Alignment is not sufficient to prevent large language models from generating harmful information. arXiv. https://arxiv.org/abs/2311.08487

[16]. Han, Y., Liu, C., & Xu, Z. (2024). Value-augmented sampling for language model alignment and personalization. arXiv. https://arxiv.org/abs/2405.06639.