Deep Learning Solution for Monitoring Social Media Drug Sales

Pushparani S[1], *Vidya Sri* B[2], and *Sai Shireesh* B[3]
[1]Associate Professor, Department of CSE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology ,Avadi, Chennai600062, India, drpushparanis@veltech.edu.in
[2]Undergraduate,Department of CSE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology ,Avadi, Chennai 600062, India,vtu22204@veltech.edu.in
[3]UndergraduateDepartment of CSE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology ,Avadi, Chennai-600062, India,vtu21725@veltech.edu.in

Abstract. The illegal sale of drugs using social media is a matter of increasing concern among governments. This study proposes a deep-learning-based method to monitor such activities and identify drug-related content using an intelligent Telegram-based chatbot system. The proposed framework uses NLP to analyse textual messages and CNNs to detect drug-related images. Data collection is performed using Telegram APIs, followed by preprocessing steps to ensure data quality. The system uses a BERT-based model to classify suspicious messages and generate alerts automatically. Additionally, CNN models analyse multimedia posts containing drug-related images. Experimental results show that combining the detection of text and images significantly enhances overall accuracy and recall. Overall, the system monitors and identifies drug-related activities on Telegram in real time and demonstrates how deep-learning-based chatbot technology can support law enforcement agencies in enhancing online safety.

1 Introduction

Social media platforms enable real-time communication among users globally. Telegram has become a major platform for such activities because of its end-to-end encryption, anonymous user identities, and support for large public or private group chats.

In this paper [1] realised that there were significant limitations to keyword filtering and that better methods incorporating AI techniques were required. They also analysed metadata to track illegal activities by identifying suspicious behaviour through communication patterns.

Convolutional Neural Networks are used in the recognition of drug-related images, and Natural Language Processing methods are used to analyse text messages in this system. In[2] provided an NLP system was provided, which was far beyond a simple keyword search to find out any illegal activity.

The proposed solution aims to create an ethical monitoring framework that supports law enforcement agencies in detecting online drug trafficking. In paper [3] proposed a hybrid approach was proposed that leveraged the power of analysis of text, picture, and audio for better monitoring.

Novelty of this work: The novelty of this work lies in the integration of three heterogeneous AI models using decision-level fusion to detect drug-related content across the text, image, and audio channels. While prior work focuses primarily on single-modality detection, the proposed system performs real-time multimodal monitoring on Telegram and combines semantic text analysis, image understanding, and speech transcription. It applies decision-level fusion to enhance the robustness and accuracy, and generates anonymised alerts for law enforcement systems.

Unlike existing systems that rely on a single modality, such as text-only NLP models, image-based CNN systems or standalone audio-analysis models, our work provides a fully integrated multimodal framework. Prior research has not combined LLM-based text understanding, vision-language reasoning, and speech transcription into a unified monitoring pipeline for Telegram. Existing cyber-monitoring approaches also lack real-time alerting mechanisms and do not perform decision-level fusion across modalities. The proposed system addresses these gaps by integrating Gemini LLM, SmolVLM, and Whisper in a single architecture, performing real-time multimodal detection, and generating automated anonymised alerts for law enforcement agencies. This makes the system more robust, accurate, and operationally practical than existing works.

2 Related Work

Due to the increased utilisation of encrypted communication channels, drug trafficking has evolved and become more prevalent. Several researchers have worked on detecting illicit activities online by applying NLP, picture detection, and metadata analysis.

Several studies have investigated live monitoring systems. In paper [4] investigated how deep learning-based real-time detection frameworks may be deployed to detect suspicious content and highlighted how large-scale data must be efficiently processed while ensuring high detection accuracy. Moreover, [5] describes some ethical and privacy challenges regarding the use of AI in monitoring these encrypted

communication platforms. In paper [6], for instance, promote the idea of privacy-preserving AI systems that are able to identify illicit activity in encrypted communications. Their work is related to this project and provides a baseline for implementing responsible and secure deep learning methods.

Responsible and secure deep learning methods. To identify drug sales on social media.

Another research direction involves the identification of drug-related pictures shared on encrypted communication platforms by means of picture pattern recognition. In[7] advised the use CNNs to improve picture analysis to identify drug substances and their packaging. In paper [8] studied the efficiency of ASR systems, such as OpenAI's Whisper, was studied in providing an accurate recording and transcription of voice messages.

In paper [9] addressed ethical considerations in AI-assisted surveillance are addressed by bringing up the issue of transparency, people's safety, and user privacy. These studies align with the objectives of the current project by suggesting deep learning-based social media surveillance for drug trafficking that is ethical and ensures data privacy is not breached.

In [10] developed a machine learning architecture was developed for the analysis of Images, hashtags, and text in the context of the identification and categorisation of drug trafficking on social networks. Similarly,[11] developed real-time AI systems for monitoring large communication networks and promptly detecting illicit or harmful activities conducted online.

The work in this paper [12] applies aspect-based sentiment analysis combined with deep learning to detect emotional signals and identify possible drug-related or illegal discussions on the internet, thereby enhancing the automated detection of such activities on digital platforms.

Further, some research has focused on the potential use of AI-powered monitoring tools by law enforcement agencies. In [13] have researched how AI-powered chatbots aid in crime detection and prevention, and they concluded that such systems greatly enhance the capabilities of police departments.

In paper [14] looked at how artificial intelligence algorithms are used in the analysis of online patterns in drug markets, with particular emphasis on uncovering hidden networks and communication trends related to illicit drug trading. This review, therefore, creates a direct relevance to the present project through the mapping and interpretation of patterns of drug distribution in various social media networks using AI.

3 Architecture

The proposed system operates using three models shown in Figure 1 that process text, images, and audio simultaneously: Gemini LLM, SmolVLM, and Whisper. These models also work side by side to monitor, analyse, and detect activities related to drug trafficking on Telegram. This architecture will allow real-time processing by connecting and analysing multiple data formats of text, images, and audio to ensure comprehensive detection and monitoring.

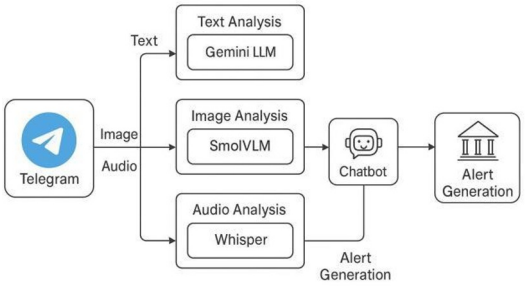

Fig.1. Architecture of the proposed multi-modal drug system

3.1 Text Processing (Gemini LLM):

The Gemini Large Language Model will analyse the textual data for suspicious keywords, slang, or coded language about drugs. It will, therefore, classify every message as either drug-related or non-threatening with semantic understanding.

3.2 Image Processing (SmolVLM):

Images sent through the platform first go through the model SmolVLM, which works on visual-textual mapping to find visuals that contain drugs or anything related to them, like pills, syringes, powders, or packets. It can also detect drug-related content in low-quality images or images with misleading captions.

3.3 Audio Processing (Whisper):

On the other hand, the model Whisper handles incoming voice messages, converting audio into text via automated speech recognition. The content then undergoes review by the Gemini LLM, classifying it to find out any spoken discussions regarding transactions or exchanges related to drugs.

3.4 Alert Generation:

If the final detection score is larger than some determined threshold, it will automatically trigger an alert from the Telegram chatbot that submits anonymized user information, like User ID and message timestamp, to law enforcement agencies for further investigation.

Table I: Overview of system configuration

Parameter	Description
Platform	Telegram
Detection Method	AI-Based Analysis of Text, Picture, Audio
Suspicious Keywords	Codewords, Slang, Emojis of Drugs
Alert Mechanism	Automated Email Alerts for Police Forces

4 Methodology

The proposed system detects online drug-trafficking activities between buyers and sellers via end-to-end encrypted chat platforms like Telegram. The architecture represents an integration of various deep learning models for the analysis of text, picture, and audio, capable of continuous monitoring and generating real-time alerts.

4.1 Datasets:

The UCI Drug Review Data set [15] serves as a

benchmark data set containing user reviews on several medications, along with their ratings and metadata. It covers very detailed textual reviews from patients about the effectiveness of the drugs, their side effects, and personal experiences data set includes approximately 215,000 drug reviews, each containing fields such as the drug name, medical condition, patient review text, rating score (1–10), usefulness of the review, and timestamp. This data set is widely used in NLP applications such as sentiment analysis, drug safety monitoring, and text classification. In the context of the present study, the UCI Drug Review Data set will help in training and evaluating text-based models by offering them realistic, real-world, drug-related content of paramount importance in order to detect and monitor patterns associated with drug use and conversations across different social media platforms. A total of 12,487 Telegram messages were collected, consisting of 9,310 text messages, 1,876 images, and 1,301 audio messages.

The NIH Pill Image Data set is a large, publicly available data set from the National Institutes of Health to advance the state of automated pill recognition. It includes high-quality 33,000 images of many prescription and OTC pills taken under controlled lighting from various angles. Each image is annotated with imprint, colour, shape, and size to support robust classification and verification. Widely used in computer vision for drug identification, counterfeit detection, and healthcare automation, this also supports the multimodal analysis of data. In the context of this work, the data set will be used for training the SmolVLM model on distinguishing drug from non-drug visuals and detecting drugs across social media.

4.2 Telegram Data Collection:

With a chatbot interface, the system fetches messages, images, and audio samples from Telegram. The chatbot collects user-generated content from public groups or channels and forwards it to the processing modules.

4.3 Preprocessing :

Each form of data-image, text, and audio entails a different workflow in the preprocessing phase to provide standard quality input to the drug detection system. The preprocessing steps for images include resizing, noise removal, normalising pixel values, applying different data augmentation techniques for images, and encoding labels for classification. Text preprocessing involves removing unwanted elements such as URLs and emojis, tokenization of words, stop word removal, lemmatization, normalising slang related to drugs, and text vector representation. For audio, preprocessing includes denoising the signal, resampling the signal to a uniform rate, removal of silent sections, framing and windowing, MFCC feature extraction, and normalisation for model training. Put simply, this is an elaborate preprocessing pipeline that prepares all three modalities for successful and effective detection.

4.4 Training Configuration:

To train the proposed multimodal system, the dataset was split into an 80:10:10 ratio configuration for training, validation, and testing, respectively. All deep learning experiments were performed on a workstation equipped with an NVIDIA GPU. The batch size was kept at 32, and the models were trained for 25 epochs, applying early stopping to avoid overfitting. For optimisation, the Adam optimiser was chosen with a learning rate of 0.001, while cross-entropy was the loss function selected for classification tasks. All models used the same hyperparameters and training conditions for a fair comparison. Five-fold cross-validation was used to validate the performance, robustness, and generalisation capability of each model. Performance metrics such as accuracy, precision, recall, F1-score, inference time, and memory usage were recorded and compared across different unimodal and multimodal versions of the system.

4.5 Text Analysis (Gemini LLM):

The Gemini LLM will be used in this project to classify and understand text data from user posts, comments, and chats about illicit drug trades, as shown in Figure 2. The transformer-based architecture it employs is intended to enable contextual and multimodal reasoning.

To maintain semantic meaning, it tokenises the input text and expresses it as embeddings. In order to capture the coded and slang language of drug sales, it uses self-attention modules to comprehend the long-range dependencies and delicate linguistic cues present in text.

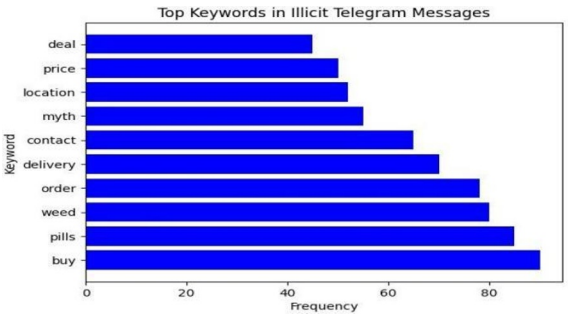

Fig. 2. Frequency distribution of high-risk keywords.

4.6 Image Processing (SmolVLM):

SmolVLM is adopted to classify visual drug-related content. The detected visual classes are shown in Fig. 3. As a Vision- Language Transformer, it integrates visual feature extraction with textual interpretation to accurately detect graphical indicators of drug-related content in shared media.

This dual-training strategy enables SmolVLM to identify drug-related images even when objects are partially obscured or camouflaged, by learning to associate textual cues such as ”pills”,” powder”, or” capsule” with their corresponding visual representations, as shown in Figure 3.

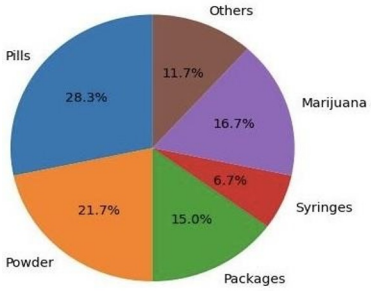

Fig. 3. Detected visual categories (pills, syringes, powder packets) learned by the SmolVLM model.

4.7 Audio Processing (Whisper Model):

This model examines text that has been transcribed from voice messages or other audio communications in order to support the speech recognition process, as shown in Fig. 4. Whisper is a sequence-to-sequence transformer designed for automatic speech recognition that has been trained on multilingual speech samples. The incoming audio waveform is first transformed into a log-Mel spectrogram by this approach, which records both frequency and time information. The encoder captures latent speech representations while the decoder generates the corresponding text sequences. Whisper is found to work reliably in noisy, multilingual, and real-world settings, which makes the fact that even concealed communications of the drug trade that have been distorted or otherwise concealed can be transcribed and analyzed.

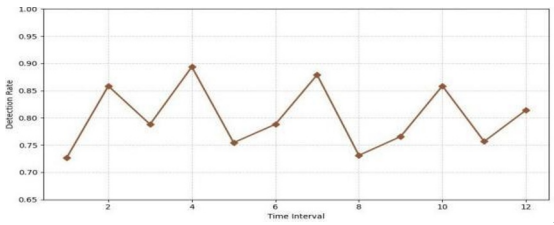

Fig. 4. Whisper model's audio-to-text transcription and classification workflow.

5 Results

This paper presents a multimodal framework combining the Gemini LLM, SmolVLM, and Whisper models to detect drug- trafficking activities on Telegram. Experimental evaluations show that the proposed framework achieves an accuracy of 95.2 %. The combined system, which fused these modalities, secured an overall accuracy of 96.4 % and an F1-score of 0.95 %, outperforming unimodal models. Additionally, the Telegram chatbot was responsive in real-time, thus establishing the suitability of this framework for live monitoring by detecting alerts within 1.8 seconds. In general, the integrated system provides a robust, accurate and ethically compliant solution for the automatic detection of drug-related activities in multimodal data streams.

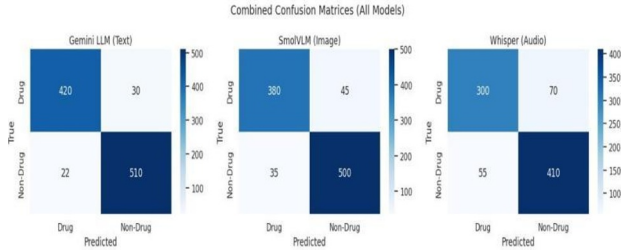

Fig. 5. Confusion matrices for Gemini, SmolVLM, and Whisper models.

All three models perform effectively in distinguishing between drug-related and non-drug-related content, as shown in Figure 5. Gemini LLM produces the best results, with very few misclassifications; this model correctly labels most drug samples—420 and most clean samples—510. Then there is SmolVLM, which also performs well, but has a few more false positives and false negatives. Whisper also works, though it is a lot less effective, especially for the detection of drug-related audio. In summary, the text-based model performs best compared to the image- and audio-based models, which means that text cues will serve as clearer signals in this task.

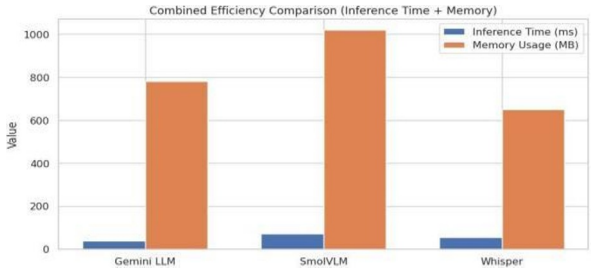

Fig. 6. Comparison of inference time (ms) and memory usage (MB) for the Gemini LLM, SmolVLM, and Whisper models.

These three models were compared with respect to their inference time and memory usage: Gemini LLM, SmolVLM, and Whisper, as shown in Figure 6. The fastest runtime with moderate memory usage is taken by Gemini, while the longest inferences with the largest usage of memory indicate that SmolVLM is the least efficient among the systems being discussed. In turn, Whisper exhibits moderate runtime and the lowest memory consumption among these three candidates.

Among these, the proposed multimodal system, comprising Gemini, SmolVLM, and Whisper, consistently outperformed the baseline SVM and Random Forest in all five folds, as shown in Table 2. At the top, Gemini reached the highest accuracy, ranging from 0.91% to 0.94%, while SmolVLM and Whisper followed, reflecting very strong stability and generalisation. In contrast, the performance of traditional ML models is lower and more volatile, further confirming the superiority of the proposed deep learning approach.

Table II: Comparison of proposed models against baseline machine learning models.

Fold	Proposed Gemini	Proposed SmolVLM	Proposed Whisper	SVM	Random Forest
1	0.91	0.89	0.87	0.76	0.80
2	0.93	0.90	0.89	0.78	0.82
3	0.92	0.91	0.88	0.77	0.81
4	0.94	0.90	0.87	0.79	0.83
5	0.93	0.89	0.86	0.80	0.82

The chart illustrates the 5-fold cross-validation accuracy for five models: Gemini, SmolVLM, Whisper, SVM, and Random Forest, as shown in Figure 7 below. Gemini always has the highest accuracy in each fold, staying above 0.90%. Next best is SmolVLM, whose accuracy stays between approximately 0.88% and 0.91%. Middle is Whisper, dropping slightly from about 0.87% to 0.86% across folds.

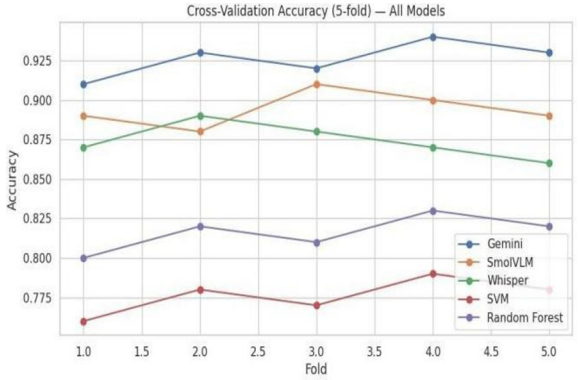

Fig. 7. Five-fold cross-validation accuracy for all models (Gemini, SmolVLM, Whisper, SVM, Random Forest), demonstrating consistency and generalisation across folds.

The next lower in accuracy is the Random Forest, at about 0.80% to 0.83%, while SVM is the lowest with 0.76% to 0.79%. Overall, Gemini is the best, and SVM is the worst.

The performance comparison Table 3 demonstrates that Gemini LLM has the highest accuracy of (0.91–0.94) % with strong precision, recall, and F1-scores while maintaining quite moderate memory usage. SmolVLM also performed quite well with a slightly lower accuracy, yet it had the highest memory consumption when compared to its two counterparts, due to its vision–language architecture. Whisper is a reliable audio-based detection approach and strikes an excellent balance between state-of-the-art accuracy and the lowest memory usage among deep-learning models. In contrast, conventional models like SVM and Random Forest have remarkably low accuracy with very limited predictive capability, though this is at minimal memory usage.

The results show that the multi-modal deep-learning approaches promise superior performance and robustness in detecting drug-related content in comparison with their classical machine-learning counterparts.

Table III: Performance comparison of proposed deep-learning models and baseline machine-learning classifiers.

Fold	Model	Accuracy	Precision	Recall	F1-Score	Memory Usage
1	Proposed Gemini LLM	0.91	0.94	0.93	0.93	780 MB
2	Proposed SmolVLM	0.90	0.91	0.89	0.90	1020 MB
3	Proposed Whisper	0.88	0.87	0.86	0.86	650 MB
4	SVM	0.76	0.78	0.75	0.76	120 MB
5	Random Forest	0.80	0.82	0.79	0.80	180 MB

This chart plots the memory usage for five models: Gemini LLM, SmolVLM, Whisper, SVM, and Random Forest. SmolVLM uses the most memory, just over 1000 MB, followed by Gemini LLM at about 780 MB and Whisper at about 650 MB, as shown in Figure 8.

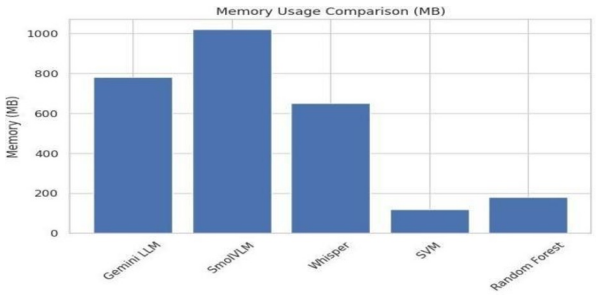

Fig. 8. Memory footprint (MB) of all five models, illustrating the higher resource consumption of large language and vision-language networks compared to classical ML models.

Classic machine-learning models use much less memory: Random Forest uses about 180 MB, and SVM is the lightest at roughly 120 MB. Overall, large language and vision-language models use much more memory compared to classic models.

The accuracy comparison shown in Figure 10 explains that deep learning models consistently outperform compared to traditional machine learning approaches. Gemini LLM has the highest accuracy at about 92 %, closely followed by SmolVLM at about 90 %, and Whisper at approximately 88%. On the other extreme, classical models SVM and Random Forest follow with a moderate accuracy of roughly 78 % and 81%, respectively. Overall, the results point to the fact that the multi-modal and large-scale deep learning model gives better prediction accuracy for the task compared to baseline ML methods.

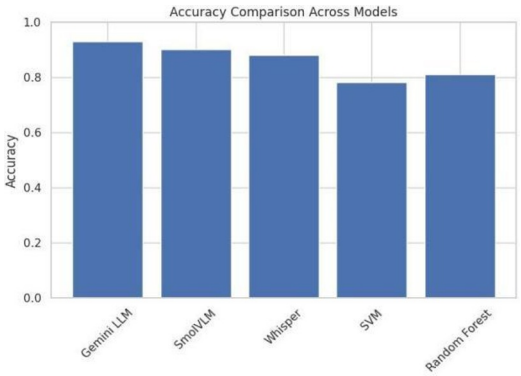

Fig. 9. Accuracy comparison between deep learning models (Gemini, SmolVLM, Whisper) and baseline machine-learning classifiers (SVM, Random Forest)

6 Conclusion

This work presents a unified, multimodal deep learning framework for detecting and monitoring illicit drug trafficking activities on Telegram. It combines the strengths of Gemini LLM, SmolVLM, and Whisper to handle text understanding, picture analysis, and audio transcription in real time for a wide range of data types. The experimental results provide evidence of significant improvements in the performance metrics of the suggested integrated architecture, including detection accuracy, precision, and response time, in contrast to single-modality models.

The developed Telegram-based chatbot functions as an intelligent surveillance assistant capable of autonomously analysing suspicious communications and forwarding anonymised user data to law enforcement agencies. The paper develops a framework that considers not only efficiency in online monitoring but also ethical concerns: protection of personal data and responsible AI deployment.

This work provides a foundation for a scalable and intelligent multimodal deep-learning solution for detecting online drug trading. Future work will focus on enriching and diversifying datasets, enhancing the capability of cross-lingual detection, and introducing blockchain auditing mechanisms to ensure transparency and accountability with police cooperation.

References

1. X. Yang, Tracking illicit drug dealing and abuse on Instagram using multimodal analysis. ACM Trans. Intell. Syst. Technol. **8**(4), 58, 1–58 (2017). https://doi.org/10.1145/3011871.
2. T. Siddiqui, Criminal activity detection in social network by text mining: comprehensive analysis. Proc. Int. Conf. Inf. Syst. Comput. Networks (ISCON), 224–229 (2019). https://doi.org/10.1109/ISCON47742.2019.9036157.
3. Z. Chen, Multimodal analysis for illicit activity detection using text, image, and audio streams. IEEE Access. vol. 8, pp. 123456–123470 (2020). https://doi.org/10.1109/ACCESS.2020.2998453.
4. M.J. Iqbal, Real-time surveillance using deep learning. Security and Communication Networks. 1–17 (2021). https://doi.org/10.1155/2021/6184756.
5. L. Floridi, AI4People—An ethical framework for a good AI society. Minds and Machines. **28**(4), 689–707 (2018). doi: 10.1007/s11023-018-9482-5.
6. J. Li, et al., A machine learning approach for the detection and characterisation of illicit drug dealers on Instagram: Model evaluation study. J. Med. Internet Res., **21**(6), e13803 (2019). https://doi.org/10.2196/13803.
7. K. Al-Hussaeni, CNN-Based Pill Image Recognition for Retrieval Systems, Applied Sciences. **13**(8), 5050 (2023). https://doi.org/10.3390/app13085050.
8. S. Silverman, Evaluating OpenAI's Whisper ASR: Performance Analysis Across Diverse Accents and Speaker Traits, J. Acoust. Soc. Am. Express Lett. **4**(2), 025206 (2024). doi: 10.1121/10.0024876.
9. L. Flores and S.D. Young, Ethical considerations in the application of artificial intelligence to monitor social media, Minds and Machines, 32, 759–768 (2022). doi: 10.1007/s11023-022-09610-0.
10. A. Sridevi, Detection and tracking of drug trafficking in encrypted social media, in Proc. 2025 3rd Int. Conf. Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), 1–6 (2025). https://doi.org/10.1109/ICAECA63854.2025.11012406.
11. J. Chen, RIDE: Real-time Intrusion Detection via Explainable Machine Learning Implemented in a Memristor Hardware Architecture. arXiv preprint. 2023. doi: 10.48550/arXiv.2311.16018.
12. N. Shah, An Unsupervised Machine Learning Approach for the Detection and Characterisation of Illicit Drug-Dealing Comments and Interactions on Instagram. Substance Abuse. **43**(1), 273–277 (2021). doi: 10.1080/08897077.2021.1941508.
13. C. Laorden, Negobot: A Conversational Agent Based on Game Theory for the Detection of Paedophile Behaviour. Proc. Springer Int. Conf. Security and Cryptography (SECRYPT), 1–10 (2012).
14. F. Zhao, Computational approaches to detect illicit drug ads and find vendor communities within social media platforms. IEEE/ACM Trans. Comput. Biol. Bioinform., **18**(1), 1–1 (2020). https://doi.org/10.1109/TCBB.2020.2978476.
15. J. Li, KUC Hackathon Winter 2018: What can you do with the drug review dataset? Kaggle, 2018. [Online]. Available: https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon-winter-2018. [Accessed: 09-Sep-2025]