International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.026 ISSN: 2581-4621

Knife Detection System Using CLIP-Based Vision-Language Model

M. Lenin Babu[1], Y. Manjula Brahmeswari[2*], N.Anusha[3], M. Vijaya Lakshmi[4],M.Srikanth[5]
[1]*Assistant Professor, Department of ECE, Bapatla Engineering College, Bapatla, India.*
[2*,3,4,5]*UG Student, Department of ECE, Bapatla Engineering College, Bapatla, India.*

***Abstract*—**Knife detection in restricted areas is essential to ensure security and preventing harm. In institutions such as schools, airports, prisons, or public transport, restricted areas can easily turn into a threat if someone carries a knife. Knifes are detected to prevent possible threats, eliminate injuries, and provide a safe environment for all. AI based Knife detection ensures safety, preventing harm, saves lives, promotes safe public spaces. In this work a machine learning algorithm known as ViT B/32 a Contrastive Language–Image Pretraining (CLIP) based model is presented for efficient detection of knifes in small spaces utilizing the Vision Language Models (VLM). CLIP is a zero-shot detection model, it can classify and interpret data without training, testing, and validation. CLIP leverages pre- trained knowledge from large-scale textual and visual data to make accurate predictions directly. This eliminates the need for the typical data preparation and segmentation process, making it highly efficient for tasks across diverse domains. It examines an image of a knife using a vision encoder to identify its shape, dimensions, and texture as patterns (feature embeddings). The proposed model has been tested with three different knife image datasets. The performance metrics considered for the simulation are precision, recall, F1 score, and accuracy. The dataset1 consists of 400 knife images, dataset2 has 4,075 knife images and dataset3 comprises 3,985 knife images. The proposed algorithm Vit - B/32 outperforms exceptional results with an average precision of 97.67%, accuracy of 89.67%, recall of 84.33%, and F1-score of 89.33% for all the three datasets. A comparison with existing algorithms highlights the effectiveness of the proposed algorithm.

***Index Terms*—Knife detection, Real-time surveillance, Threat alert, Image processing, Preventive policing.**

I. INTRODUCTION

Knife detection [1] refers to the process of identifying the presence of knife, typically using technology like computer vision models (eg: CLIP). It is usually used in secure environments like airports, schools and public areas to detect possible threats and provide security. Advanced systems are capable of detecting knives by analyzing visual patterns or materials, enabling quick and efficient identification. Knife detection [2]is a critical component of modern security measures, ensuring safety in various settings by identifying knives promptly. These systems help prevent violent incidents in public spaces and enable swift responses to potential threats. In workplaces such as hospitals, offices, and correctional facilities, knife detection fosters a secure environment for staff and visitors, minimizing risks of harm or violence. Additionally, these systems act as a deterrent for individuals in carrying weapons into restricted areas, thus aiding crime prevention. Leveraging advanced technologies like artificial intelligence and machine learning enhance their efficiency and precision through real-time monitoring and automatic alarms. Their adaptability allows for tailored solutions addressing specific safety concerns, from mass events to high-threat environments like prisons and courthouses. By promoting a culture of safety and reassurance, knife detection not only protects lives but also fosters peace of mind in communities. The ViT-B/32 [3] CLIP (Contrastive Language–Image Pretraining) model, developed by OpenAI, is a powerful AI tool well- suited for knife detection. CLIP bridges the gap between images and text, excelling in object recognition tasks. In knife detection, it associates knife images with textual descriptions, eliminating the need for

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.026 ISSN: 2581-4621

traditional training methods. CLIP can analyze real-time video feeds, identifying potential threats even in challenging conditions such as poor lighting, partial visibility, or unconventional angles. Its versatility enables it to generalize across datasets, making it effective for varied environments like airports, schools, or workplaces. Fine-tuning CLIP with custom datasets enhances its accuracy for specific applications. By learning a shared representation of images and text, CLIP recognizes knives based on their visual and descriptive features, distinguishing them from similar objects. This capability allows it to perform precise detection in static images or dynamic video feeds. When integrated into surveil- lance systems, CLIP enhances real-time monitoring, triggering alerts in high-risk areas such as schools, transport hubs, or event venues.

Its adaptability extends to forensic applications, aiding law enforcement by analyzing crime scene footage to classify knives based on their characteristics. CLIP's ability to generalize across diverse datasets and recognize objects not explicitly trained on makes it invaluable for real-world knife detection scenarios. Its combination of precision, adaptability, and versatility significantly contributes to public safety and crime prevention. Globally [4], Knife detection systems have diverse applications that contribute to enhancing safety and security in various environments. They are used in public spaces such as airports, train stations, and shopping malls to identify concealed weapons and prevent potential threats. These systems play a vital role in securing large events like concerts and sports gatherings, ensuring the safety of attendees. In high-risk workplaces, such as hospitals, correctional facilities, and government offices, knife detection safeguards employees and visitors from harm. Law enforcement agencies leverage these tools to proactively detect weapons and respond to threats in high-crime areas. Educational institutions also utilize knife detection systems to protect students and staff, particularly in regions with a history of violence. Additionally, transportation hubs such as bus terminals, subways, and air- ports rely on these systems to maintain passenger safety. Retail and commercial spaces benefit from knife detection technologies by deterring theft and providing secure shopping environments. Furthermore, AI-powered knife detection integrated with smart surveillance systems ensures real-time monitoring and automated alerts, making these solutions indispensable for creating safer communities. Knife-related [5], incidents globally highlight the critical need for advanced detection systems to enhance safety and security. In 2018, the United States reported over 123,000 aggravated assaults involving knives or cutting instruments, and similar trends have been observed in the United Kingdom, where rising knife crime rates have led to stricter laws and increased security measures. Public spaces like airports, schools, and event venues have benefited significantly from the implementation of knife detection systems, which have contributed to a noticeable reduction in violent incidents and fostered safer environments. Urban areas, in particular, face higher rates of knife-related crimes due to their dense populations and complex social dynamics. Technological advancements, such as the integration of the CLIP model and real-time surveillance systems, have revolutionized knife detection by enabling quick and precise identification of threats, even in challenging conditions. However, implementing these systems comes with challenges, including high costs, privacy concerns, and occasional false positives. Despite these hurdles, the refinement of AI technologies and collaborative efforts between governments, tech developers, and communities have the potential to make knife detection systems more accessible and effective, ensuring public safety while addressing ethical considerations.

II. LITERATURE SURVEY

In the past, Numerous studies on knife detection have been conducted, often as part of research publications and technical reports focusing on security and surveillance applications. These studies explore various approaches, including computer vision, deep learning, and infrared imaging, to accurately identify and classify knives in real-world scenarios.

M. Milagro Fernandez-Carrobles [6] studied a system that detects guns and knives in video to improve safety in public spaces. It uses Faster R-CNN with GoogleNet and SqueezeNet for image

DOI:10.46647/ijetms.2025.v09i02.026 ISSN: 2581-4621

analysis. Tests showed SqueezeNet was best for guns, with 85.44% accuracy, while GoogleNet was better for knives, with 46.68% accuracy. Although these results are better than past studies, knife detection is still less accurate compared to guns.

Mai K. Galab [7] introduced a method to detect knives in surveillance videos using deep learning. Knives are hard to spot because of their different shapes, sizes, and reflective surfaces. The method measures frame brightness and uses Gamma Correction to improve visibility in bright frames. Among various models tested, AlexNet worked the best, achieving 96.95% accuracy even in poor lighting. This system is useful for real-time safety, like in CCTV setups. However, relying only on Gamma Correction may not handle tricky lighting, such as shadows or sudden light changes. Using more advanced techniques or extra sensors could make it better.

Muhammad Tahir Bhatti [8] developed a system to detect weapons, like pistols and revolvers, in CCTV footage using deep learning. The goal was to automate weapon detection and reduce human monitoring. Challenges included different angles, blocked views, and a lack of standard datasets. Among tested models, YOLOv4 performed best with 91% accuracy. Techniques like grayscale were used to improve detection in low light. The system worked well in dark and low-resolution settings, making it suitable for public safety. However, it needs more testing in challenging conditions like extreme lighting, crowded areas, and unusual camera angles to improve accuracy for real-world use.

Yoshemart Amador-Salgado [9] introduced a system to detect knives in indoor CCTV footage. It addresses challenges like light reflections on metallic knives, varying distances, and low-resolution images. The system uses techniques such as edge detection and contour analysis to find knives in video frames. It works within 0.5 to 4 meters and detects one knife per second under white or infrared lighting. This real-time system is useful in places like stores and banks, improving safety by fitting into existing setups. However, it requires knives to be at least 10 cm long and 4 cm wide to work well, and lighting reflections can still cause issues. More testing is needed for objects that look like knives and diverse conditions to ensure better performance.

Daniele Berardini [10] developed a system to detect hand- guns and knives in indoor CCTV videos using deep learning, designed for low-cost devices like the NVIDIA Jetson Nano. The system works in two steps: one CNN detects people in video frames, and another CNN looks for weapons in those areas. It performed well with an average precision of 79.30% and processed 5.10 frames per second. Useful for places like hospitals and airports to improve safety. However, the frame rate may not be fast enough for busy or dynamic situations, and it is only tested for indoor use, making it less reliable for outdoor conditions with changing weather or lighting.

Mr. S. Barath [11] proposed a system to detect weapons like guns in CCTV footage using deep learning and OpenCV. It aims to improve security by automating weapon detection, which is done through methods like grayscale conversion, edge detection, and object classification. The system also sends email alerts to authorities when a weapon is found. It is useful in places like malls and cinemas but faces challenges such as false alarms, struggles with small or hidden weapons, and reliance on good-quality footage. Improving the dataset and using more advanced algorithms could make it more reliable and adaptable to real-world conditions.

Eungyeom Ha [12] introduced a dataset to improve harmful object detection for online platforms and surveillance. It includes both easy and hard cases, such as small or hidden objects, to challenge detection models. Models like YOLOv5 and Faster R-CNN showed better results when trained on this dataset, especially with difficult cases. The dataset, codes, and trained models are freely available to help create advanced systems for tasks like content moderation and real-time security.

III. DATASET DESCRIPTION

Data has an important contribution towards the learning of any Machine learning model as the model learns and derives feature from it. The knife dataset is drawn from [16]. The dataset1 contains 400 training images, and their resolutions are resized to 100x100 pixels. Another knife

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.026 ISSN: 2581-4621

dataset is collected from [17] which Comprises of total of 4,075 images designed for object detection tasks. This structured division allows for machine learning models in projects focused on knife detection. Another knife dataset3 is extracted from [18], It contains a total of 4,075 images. Each image has been pre-processed to a size of 416x416 pixels.

Fig. 1. Dataset for proposed methodology

IV. PROPOSED METHODOLOGY

The CLIP (Contrastive Language – Image Pretraining) model, developed by OpenAI, aligns visual and textual in- formation, making it a versatile tool for various tasks, including knife detection. To implement CLIP for this purpose, a dataset containing images of knives and non-knife objects must be collected and preprocessed, with images resized and labelled appropriately. Although CLIP can function without modification, fine-tuning it on this specific dataset enhances its accuracy in detecting knives. The model, once trained, analyzes input images to identify the presence of knives based on its visual-textual alignment capability. Finally, this trained model can be integrated into systems such as surveillance applications for real-time knife detection, which is particularly useful for security purposes. Proper care should be taken to ensure the model's accuracy to minimize errors and deploy it responsibly. Architecture of the CLIP model is illustrated in the below Fig 2.

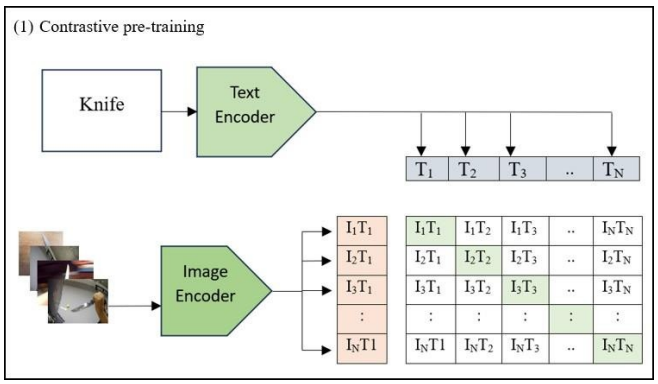
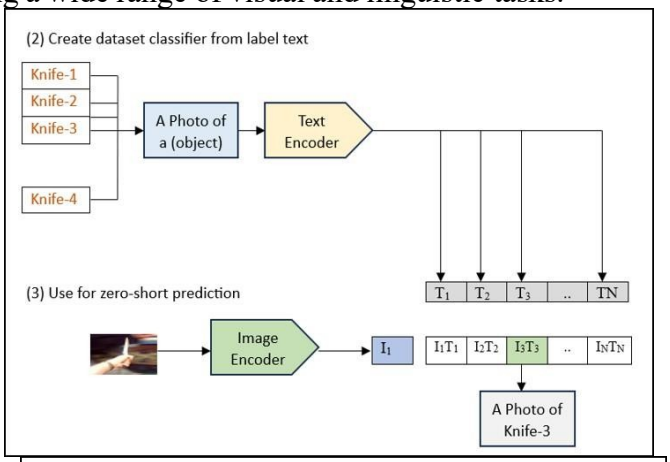


Fig. 2. Contrastive Pre - Training

A Contrastive pre-training involves three main stages: contrastive pre-training, where both text and image encoders map their inputs into a shared embedding space; classifier creation, where text descriptions or labels are encoded to serve as classification references; and zero-shot prediction, where an image is encoded and compared with the text embeddings to determine the best match. This method allows the model to interpret and classify images without requiring specific training for each task. It starts with contrastive pre- training, where the model's text and image encoders work

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.026 ISSN: 2581-4621

together to map text descriptions and images into a shared embedding space. This alignment ensures that semantically related text and image embeddings are positioned closer together in the shared space, making meaningful comparisons possible. Next, the process moves to dataset classifier creation, where predefined text labels or phrases, such as "knife" or "tree," are encoded using the text encoder to generate embeddings that serve as classification references. These text embeddings essentially become benchmarks against which new image embeddings are compared. In the final phase, zero- shot prediction, the image encoder processes a new image to create its embedding, which is then compared with the label embeddings using a similarity metric, such as cosine similarity. The model identifies the label with the highest similarity score as the best match for the image. This structured process demonstrates the model's ability to classify or identify objects in images without requiring additional task-specific training, highlighting its adaptability and efficiency in handling a wide range of visual and linguistic tasks.



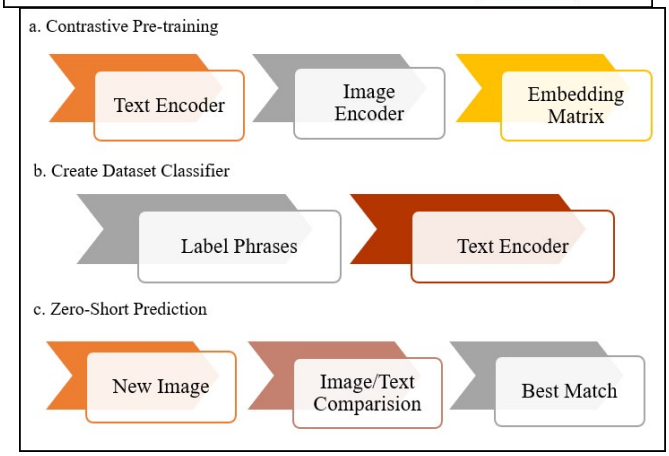


Fig. 3. Zero-short prediction
Fig. 4. Flowchart CLIP Model

V. PROPOSED ALGORITHM FOR CLIP MODEL

The flowchart outlines a systematic and well-defined process that incorporates the principles of contrastive learning, the creation of a dataset classifier, and the application of zero- shot prediction methodologies. It begins with the phase of contrastive pre-training, during which a Text Encoder and an Image Encoder are collaboratively trained. These two components work in tandem to map both textual and visual inputs into a unified Embedding Matrix, a shared representational space. This alignment ensures that semantically related text and image representations are closely positioned within this matrix, facilitating meaningful

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.026 ISSN: 2581-4621

VI. RESULTS AND SIMULATION

All Results and simulation are done by using the Google Colab environment and the Pytorch framework for machine learning tasks, specifically integrating the Hugging Face Trans- formers library to use the ViT-B/32 CLIP (Contrastive Language–Image Pretraining) model developed by OpenAI. In a detection task there are two possible results, positive and negative. Some positive cases can be classified as negative and vice versa. These cases are called false positives and false negatives, respectively. Thus, the following four cases are considered: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

Evaluation Metrics:

$\frac{TP + TN}{}$ .....

comparisons. Following this, the process transitions to the dataset classifier creation phase. At this stage, predefined Label

$\text{Accuracy} = TP + TN + FP + FN$ (1)

Phrases, which consist of textual descriptions representing possible classification categories, are processed through the Text Encoder. This step generates corresponding embeddings

$\text{Precision} = \frac{TP}{TP + FP}$ ..... (2)

for these labels, effectively creating a set of reference points that serve as anchors for subsequent classification tasks. These embeddings act as a textual blueprint for the model to classify various inputs. The next step in the process involves zero-shot

$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$ ..... (3)

prediction, a technique designed to classify new data without requiring prior training on that specific dataset. During this phase, a New Image is introduced and passed through the Image Encoder. The encoder transforms the image into its corresponding embedding, which resides in the same shared representational space as the text embeddings. An Image/Text Comparison mechanism is then employed to analyze and compare the newly generated image embedding with the precomputed text embeddings. Finally, in the best match selection stage, the model evaluates the degree of similarity between the image embedding and each of the text embeddings. The model ultimately identifies the text embedding that most closely aligns with the image embedding, determining the most relevant category. The Flowchart for the CLIP model is given below fig 4 .

$\text{Recall} = \frac{TP}{TP + FN}$ ..... (4)

TABLE I

PERFORMANCE METRICS FOR DATASET-1

Category	Metrics	Value (%)
Performance Metrics	Precision	95.02
	Accuracy	98.42
	Recall	95.5
	F1-Score	95.26

In the model showed 95.02% precision, 98.42% accuracy, 95.5% recall, and an F1-score of 95.26%. It per- formed well overall, with balanced precision and recall shown in table 1. In the model achieved 99.18% precision, 72.56% accuracy, 59.63% recall, and an F1-score of 74.47%. It was accurate but missed some knife instances. shown in table 2

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.026 ISSN: 2581-4621

TABLE II
PERFORMANCE METRICS FOR DATASET-2

Category	Metrics	Value (%)
Performance Metrix	Precision	99.18
	Accuracy	72.56
	Recall	59.63
	F1-Score	74.47

The system excels with balanced and high metrics : precision (99.41%), accuracy (98.62%), recall (98.4%), and F1-score (98.9%), making it the best among the three. shown in table 3

TABLE III
PERFORMANCE METRICS FOR DATASET-3

Category	Metrics	Value (%)
Performance Metrix	Precision	99.41
	Accuracy	98.62
	Recall	98.4
	F1-Score	98.9

The performance of three datasets across four key metrics : Accuracy, Precision, F1-Score, and Recall are shown in Figure
5. The X-axis represents these performance metrics, while the Y-axis displays the corresponding values, likely in percentage form. Each dataset is represented by a different color: blue for Dataset 1, orange for Dataset 2, and gray for Dataset-3. Dataset 1 also performs well, particularly in Accuracy (98.42) and Recall (95.26), but slightly trails Dataset 3. In contrast, Dataset 2 shows the weakest performance in Accuracy (78.56), F1-Score (59.63), and Recall (74.47), though it achieves the highest Precision at 99.18. Dataset-3 consistently outperforms the others across all metrics, achieving the highest Accuracy (98.62), Precision (98.62), F1-Score (98.9), and
Recall (98.4).

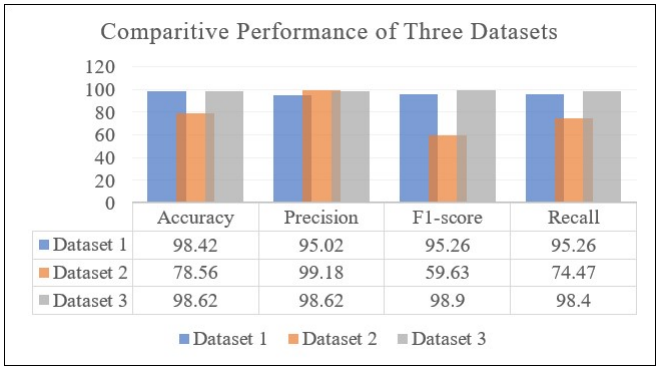


Fig. 5. Knife Detection using ViT-B/32 model Performance

Confusion Metrics
Confusion Metrics A confusion matrix is a tool used to evaluate the performance of classification models by comparing actual values with predicted values. The first confusion matrix represents a balanced model with 382 true positives (TP) and only 18 false negatives (FN), meaning it correctly identifies most positive cases while missing very few. It also has a low false positive (FP) count of

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.026 ISSN: 2581-4621

20 and 1980 true negatives (TN). shown in fig 6

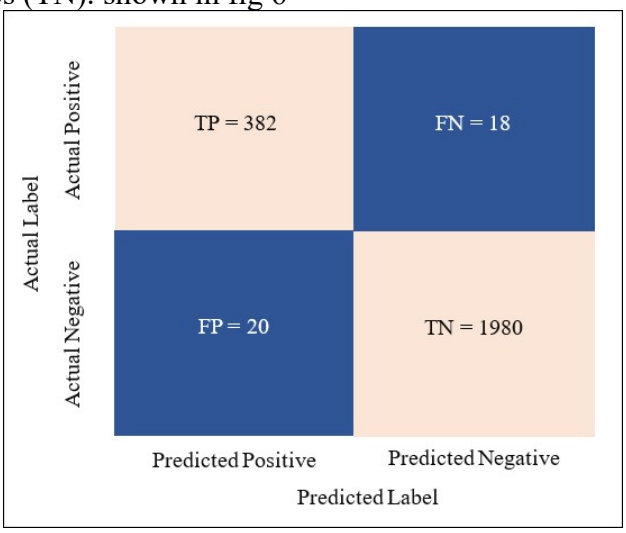


Fig. 6. Confusion Metrics For Dataset-1

In the second matrics, It includes True Positives (TP) with 2430 correctly identified positive cases, and True Negatives (TN) with 1980 accurate predictions of negative cases. Additionally, the matrix highlights False Negatives (FN) at 1645, representing missed positive instances, and False Positives (FP) at 20. Shown in fig 7.

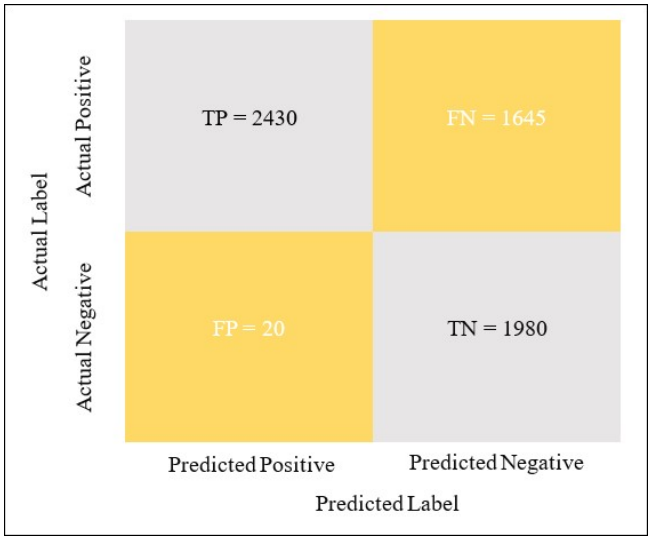


Fig. 7. Confusion Metrics for Dataset-2

In the third matrics It includes True Positives (TP), which are 3391 correctly identified positive cases, and True Negatives (TN), representing 1980 accurately predicted negative cases. Additionally, it highlights False Negatives (FN) at 55, indicating positive cases that were missed, and False Positives (FP) at 20 shown in fig 8

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.026 ISSN: 2581-4621

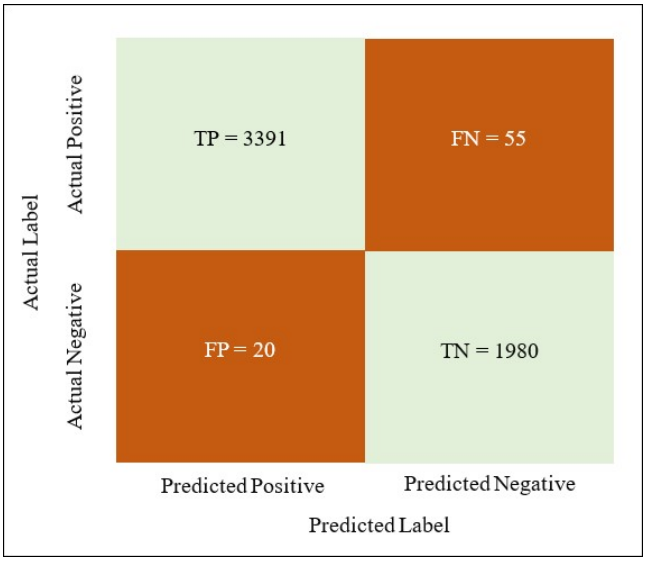


Fig. 8. Confusion Metrics for Dataset-3

Confusion matrices illuminate predictive accuracy, aiding in clear evaluation, refining algorithms, and fostering informed decisions for improved outcomes.

VII. CONCLUSION AND FUTURE WORK

The proposed knife detection system using the CLIP models utilized, specifically the ViT-B/32 model shows great results in keeping confined spaces safe. It detects knives effectively by analyzing images and matching them with text descriptions. Tests on three datasets show high performance, with accuracy, precision, recall, and F1 scores reaching up to 98%. This proves the system is reliable and better than existing methods. Future work could integrate Openpose, A human pose estimation framework, to further enhance the detection system by analyzing the context of knife usage or interaction. Openpose can enable the algorithm to detect suspicious human movements or postures associated with potential threats, thus adding an additional layer of situational awareness. This extension would strengthen the system's capability to proactively address safety concerns in environments like schools, airports, prisons, and public transport. By combining knife detection with pose recognition, the framework has the potential to become even more robust in fostering secure spaces for society.

REFERENCES

[1] Scylla Technologies Inc., "Knife Detection System," 2025. [Online]. Available: https://www.scylla.ai/knife-detection-system. [Accessed: March 21, 2025].

[2] M. Moran, A. Conci, and A´ . Sa´nchez, "Automatic Detection of Knives in Complex Scenes," in *ICT Applications for Smart Cities*, A. D. Sappa, Ed. Cham: Springer, 2022, pp. 57–77.

[3] D. A. Noever and S. E. Miller, "Knife and Threat Detectors," arXiv, 2020. [Online]. Available: https://www.academia.edu. [Accessed: March 21, 2025].

[4] D. A. Noever and S. E. Miller, "Knife and Threat Detectors," arXiv, 2020. [Online]. Available: https://www.academia.edu. [Accessed: March 21, 2025].

[5] A. Glowacz, M. Kmiec´, and A. Dziech, "Visual detection of knives in security applications using Active Appearance Models," *Multimedia Tools and Applications*, vol. 74, pp. 4253–4267, 2015.

DOI:10.46647/ijetms.2025.v09i02.026 ISSN: 2581-4621

[6] M. M. Fernandez-Carrobles, O. Deniz, and F. Maroto, "Gun and Knife Detection Based on Faster R-CNN for Video Surveillance," in *IbPRIA 2019, LNCS 11868*, Springer Nature Switzerland AG, 2019, pp. 441– 452. DOI: 10.1007/978-3-030-31321-0 38._

[7] M. K. Galab, A. Taha, and H. H. Zayed, "Adaptive Technique for Bright- ness Enhancement of Automated Knife Detection in Surveillance Video with Deep Learning," *Arabian Journal for Science and Engineering*, 2021. DOI: 10.1007/s13369-021-05401-4.

[8] M. T. Bhatti, M. Aslam, M. G. Khan, and M. J. Fiaz, "Weapon Detection in Real-Time CCTV Videos Using Deep Learning," *IEEE Access*, vol. 9,
pp. 34366–34381, 2021. DOI: 10.1109/ACCESS.2021.3059170.

[9] Y. Amador-Salgado, J. A. Padilla-Medina, M. A. Rodr´ıguez-Licea, F.
J. Pe´rez-Pinal, J. J. Mart´ınez-Nolasco, and A. I. Barranco-Gutie´rrez, "Knife Detection Using Indoor Surveillance Camera," in *Proceedings of the IEEE Artificial Intelligence of Things (AIIoT)*, 2021. DOI: 10.1109/AIIoT52608.2021.9454246.

[10] D. Berardini, L. Migliorelli, A. Galdelli, E. Frontoni, A. Mancini, and
S. Moccia, "A deep learning framework running on edge devices for handgun and knife detection from indoor video-surveillance cameras," *Multimedia Tools and Applications*, vol. 83, pp. 19109–19127, 2024. DOI: 10.1007/s11042-023-16231-x.

[11] S. Barath and S. Kanagathara, "Real-Time Weapon Detection and No- tification Using Deep Learning for Security Application," *International Journal of Research Publication and Reviews*, vol. 4, no. 6, pp. 3059– 3066, 2023. [Online]. Available: www.ijrpr.com.

[12] E. Ha, H. Kim, and D. Na, "HOD: New Harmful Object Detection Benchmarks for Robust Surveillance," in *Proceedings of the Winter Conference on Applications of Computer Vision (WACV) Workshop*, 2024. [Online]. Available: https://github.com/poori- nuna/HODBenchmark-Dataset.

[13] S. Shekhar,"Knife Dataset," 2018. [Online]. Avail- able: https://www.kaggle.com/datasets/shank885/knife- dataset?resource=download.

[14] Roboflow, "Knife Dataset - New," 2025. [Online]. Available: https://universe.roboflow.com/workspace-zqssx/knife-dataset-new. [Ac- cessed: March 19, 2025].

[15] Roboflow, "Knife Dataset - New," 2025. [Online]. Available: https://universe.roboflow.com/workspace-zqssx/knife-dataset- new/dataset/1. [Accessed: March 19, 2025].