Real-Time Multi-Modal Threat Detection: Integrating YOLOv11 and EfficientNetV2 using an Adaptive Frontend Framework

1st Aryan Dani
Dept. of Polytechnic
Dr. Vishwanath Karad MIT World Peace University
Pune, India
daniaryan212@gmail.com

2nd Prakhar Jaiswal
Dept. of Polytechnic
Dr. Vishwanath Karad MIT World Peace University
Pune, India
jas.prakhar@gmail.com

3rd M. Sobaan Jagirdar
Dept. of Polytechnic
Dr. Vishwanath Karad MIT World Peace University
Pune, India
sobaanjagirdar008@gmail.com
@gmail.com

4th Swayamprakash Patro
Dept. of Polytechnic
Dr. Vishwanath Karad MIT World Peace University
Pune, India
swayamprakashpatro@gmail.com

5th Jyoti Mante
Dept. of Polytechnic
Dr. Vishwanath Karad MIT World Peace University
Pune, India
jyoti.khurpade@mitwpu.edu.in

***Abstract*—The modern world requires surveillance and monitoring systems, yet current systems face problems such as low accuracy in threat detection and high false alarm rates, limiting their effectiveness. Additionally, human operators face cognitive overload and fatigue from constant monitoring, further reducing system reliability. We propose an automated system using an adaptive framework that selects the optimal model based on the input data type and then identifies threats and triggers an alarm system, which decreases the frequency of such attacks and relieves humans from such a tedious task, thereby improving security and public safety. The first system uses YOLOv11, a strong model highly respected for its processing speed and accuracy in detecting objects. For the second system, the model we use is EfficientNetV2, a type of CNN recognized for its efficiency, making it a go-to classifier for our use case. We have trained and tested both systems on a large dataset of knives, pistols, guns and various X-ray scans with weapon and non-weapon data. Our systems got a mean Average Precision (mAP) of 0.83 for weapon detection in real-time scenarios and 99.2% accuracy in classifying X-ray scans, significantly outperforming existing benchmarks. We have also generated synthetic data using GANs to improve the model's capability to Simplify across different scenarios. We evaluate the model's performance against metrics such as accuracy, precision, recall, F1-score, and mean Average Precision (mAP) at various intersection-over-union (IoU) thresholds, demonstrating a high capability to distinguish weapon and non-weapon classes with exceptionally low error rates.**

***Index Terms*—threat detection, automated surveillance, Yolov11, computer vision, machine learning, EfficientNetV2, deep learning, image classification, public safety, angular**

I. Introduction

Detecting threats in real time is a challenging task due to the substantial risks that undetected dangers pose to public safety. In recent times, statistics indicate an increase in violent attacks with weapons posing public safety issues [1]. With the average rate of reported attacks growing year by year, human observation and regular cameras cannot cope with the requirements of real-time situations [2]–[5]. In a world where violent attacks can happen suddenly, it is critical to identify threats in a timely manner [6]. The difficulty of detecting threats increases in densely populated public areas, where the crossing paths of individuals and objects can conceal potential threats, usually resulting in delayed or missed detections. As the number and complexity of violent incidents increase, the demand for advanced surveillance systems becomes more pressing. However, current systems often fall short of these increased demands, increasing the chance of undetected threats [7]. This disparity between increased demands and existing capabilities underscores the need for new solutions that can provide accuracy and speed in real-world applications. It becomes challenging to detect threats when multiple objects and individuals intersect in crowded places. Recent object detection models such as YOLOv11, a newer member of the YOLO family, which employs the C2PSA module for enhanced detection in crowded scenes [8], and classifiers such as EfficientNetV2, a lightweight CNN [9], are some of the most promising solutions to such dilemmas. These technologies offer significant potential to overcome the limitations of traditional surveillance, paving the way for more reliable threat detection systems. In this paper, we propose a novel real-time threat detection framework that integrates the strengths of YOLOv11 and EfficientNetV2 to address these challenges. Our approach combines the superior object detection capabilities of YOLOv11 in complex environments with the efficient classification prowess of EfficientNetV2, with the goal of achieving high precision and reduced latency in classifying threats trained on the GDXray dataset [10].

II. Literature Survey

A. Weapon Detection using Deep Learning Techniques

There are different approaches to consider when searching for a model or technique to use in a situation where the detection of threats is performed in real time. The research article [11] demonstrates various use cases of deep learning techniques/models that are commonly used for the detection of weapons. It showcases an in-depth analysis of the most common object detection frameworks, such as Convolutional Neural Networks (CNNs), YOLO (You Only Look Once), Fast R-CNN Faster R-CNN (Region-Based Convolutional Neural Network), SSD (Single Shot MultiBox Detector), and RetinaNet, and mentions their advantages and drawbacks in the detection of weapons. These models are trained and evaluated on publicly available datasets such as COCO (Common Object in Context), PASCAL VOC, and domain-specific datasets including RWD (Real World Weapon Dataset). Another research paper [7] talks about how they implemented SSD for detecting several weapons in a video frame. It had a unique approach to detecting weapons and had multiple categories to cater to. This system was tested and developed based on TensorFlow and evaluated with a 294 second long video, which showcased 7 different weapons within 5 categories including handguns and shotguns. The SSD model achieved a precision of 0.8524 and 0.7006 at two IoU thresholds (0.50 and 0.75). The paper [12] reviews the advancements in deep learning techniques applied to object detection Specifically, this paper discusses many of the renowned architectures that transformed object detection just as described above. R-CNN initially proposed the region proposal concept using CNNs for object detection. Fast R-CNN and Faster R-CNN have also greatly improved by the addition of the region proposal network (RPN) to the detection pipeline, which significantly improves speed and accuracy. Models such as YOLO and SSD are primarily used for the detection of objects in real-time, these models are precise and do not slow down significantly since they predict class probabilities and bounding boxes simultaneously with a balance between speed and accuracy, making it suitable for real-time applications. Now that we've understood that YOLO is the most suited for our use case, [6] is a research paper that employs YOLOv8 to detect firearms and other weapons in various environments achieving an mAP of 0.78 with an F1 score of 0.82. This is the most recent implementation of our use case. The model used in this research paper seems to be performing very well in real world scenarios. YOLOv11 is built on the foundation of YOLOv8 and it outperforms all the models in the YOLOv8 family by a significant margin. [8] is a research paper that proves these claims. This paper compares YOLOv11 to its predecessors on different parameters like speed and accuracy. In this paper, YOLOv11-m ranked 1st in accuracy and outperformed the entire YOLOv8 family with good enough speed making it the perfect model for our use case.

B. Classification using Deep Learning Techniques

For our second system, the main priorities we had to take into consideration were accuracy and speed with the bare minimum computational power. Searching for the best model for our use case was the first step. Liu, Dingming et al. [13] compared the results of different CNN models to predict the chance of cancer by taking into consideration 9,109 microscopic images, and the findings indicate that EfficientNetV2-b0 - b2 performs at the same level as some of the heavier models like MobilenetV2 and InceptionResnetV2 while utilizing less resources. Another research paper [14] compared different models such as InceptionNetV2, ResNet and InceptionResnetV2 for our exact same use case. In this paper, InceptionResnet when used with Faster-RCNN came out to be the most accurate with an exceptional mAP of 0.9410 but with the obvious downside of speed. The time taken to perform classification on one scan was very high for a real-world scenario, thus making this system computationally inefficient for our use case. After reading the paper [15], which compares various CNN's and NET models and several different methods to arrive at a conclusion that EfficientNetV2 paired with a vision transformer gave the highest accuracy for Breast Cancer images. The second key inference that we found was that all the EfficientNet models were the best performing models for binary classification problems like these, or even a multi classification problem at all magnification levels. Another paper [9] showcases that EfficientNetV2-s when used as a Classifier for Predicting Acute Lymphoblastic Leukemia produced an F1 Score of 97.34%. The paper by Morris et al. (2018) [16] demonstrates the application of Convolutional Neural Networks (CNNs) for detecting explosives in X-ray baggage scans, achieving a promising AUC of 0.95 using state-of-the-art models like VGG19 and InceptionV3. It introduces the Passenger Baggage Object Database (PBOD), a novel non-proprietary dataset, enabling further research in automated threat detection. Additionally, the study explores threat localization through heatmaps, highlighting the potential of CNNs to enhance security screening efficiency. Having read the paper [6], was the inspiration to use a GAN to diversify our dataset. Our research focuses on why we need synthetic data generated by GANs to offset the limited data we have for weapons. Besides these findings, our research is also an extension of work conducted by earlier research. By adding synthetic data created with the use of GANs, we were able to address the weakness of an imbalanced and small dataset for the detection of weapons. Based on our experiments, the addition of samples created by GANs not only improves the diversity of the training set but also the strength and generalization of the classifier. This combined with the new architectures such as YOLOv11 with its new architecture and EfficientNetV2 with a vision transformer has resulted in improved accuracy and detection reliability in our project.

C. Summary Table of Key Models and Performance Metrics

The Table I summarizes all the metrics mentioned in the research papers cited in the sections above

TABLE I
COMPARISON OF KEY MODELS AND THEIR PERFORMANCE METRICS FROM LITERATURE

Model	Publ.	Use Case	Key Inferences
YOLOv8 [6]	arXiv	Weapon Det.	mAP 0.78, F1 0.82
SSD [7]	IEEE	Weapon Det.	Prec. 0.85/0.70 IoU 0.5/0.75
YOLOv11 [8]	arXiv	Metric Eval.	YOLOv11-m top acc.
EffNetV2-s [9]	MDPI	Leuk. Pred.	F1 97.34%
EffNetV2-b0-b2 [13]	IEEE	Cancer Det.	Comp. to heavy, low res.
IncResV2 [14]	arXiv	Cancer Det.	mAP 0.94, slow
EffNetV2+VT [15]	IEEE	Cancer Det.	High acc. breast ca.
HVGG19 [16]	IEEE	X-Ray Threat	AUC 0.95

III. PROPOSED METHODOLOGY

In our suggested system for real-time threat detection in dense public areas, we employ YOLOv11, a newer member of the Ultralytics family. YOLOv11 accommodates several applications, such as "object detection, image segmentation classification, pose estimation, and oriented object detection" (OBB) [8]. YOLOv11 achieves these advancements with its new C2PSA module which combines cross-stage partial networks with self-attention mechanisms. In our experiments, the model worked efficiently in crowded scenes. YOLOv11 is extremely strong in situations of overlapping objects, excelling in detection and making it extremely effective in dense transit areas or event centers. Its nano-to-large, scaled architecture offers hardware deployment flexibility across edge devices and servers. We have trained it on synthetic datasets like GAN-generated data and domain-specific datasets like OD-weapon which makes it reliable for weapon detection in dynamic scenarios. To solve security and screening problems, the second system employs the EfficientNetV2 model, a convolutional neural network model developed by Google. This model is used since it is designed to be computationally efficient and perform fast inference while maintaining a high accuracy [9]. In the tests we conducted the model proves to perform better in situations where a lot of feature extraction is needed, such as the detection of concealed weapons in bags. For training the system we use the GDXray dataset [10] because it is well recognized among researchers for its accurate annotation of weapons, electronics, and background images. The annotations in this dataset are mostly used for applications aimed at deployment in airports, border security areas and areas subject to strict security legislation. To provide a comprehensive solution to this problem, we have designed an adaptive front-end application using Angular. This application integrates the real-time CCTV analysis and X-ray screening into one system for both of our deep learning models. We directly apply YOLOv11 to CCTV camera feeds to facilitate immediate threat detection while we process images from X-ray scans with EfficientNetV2 for detailed analysis and classification.

A. System Architecture

The system architecture consists of three main components:

- **Data Ingestion Layer:** Captures real-time CCTV footage and X-ray scans.
- **Processing Layer:** Runs YOLOv11 for object localization and detection on CCTV footage and EfficientNetV2 for classification on X-ray scans.
- **Frontend Layer:** An Angular-based application that displays detection results, classifications, shows analytics based on past threats and generates reports.

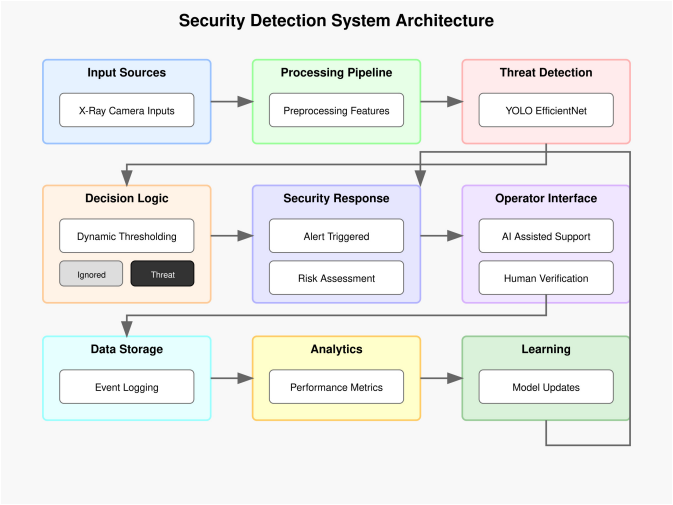

Fig. 1. System Architecture

B. Datasets Used

1) YOLOv11 Dataset: We utilized various datasets, including the OD weapon detection dataset from GitHub, containing 5,000 images of the class 'knives,' consisting of various types of knives in different enviorments and positions to stimulate a real world situation and multiple datasets from Roboflow consisting of approximately 12,000 images of 'pistols' within cluttered and occluded enviorments and 2,000 images of 'guns' that were all manually annotated on Roboflow. To enhance diversity, we applied augmentations such as brightness adjustments ($\pm 20\%$), random noise (Gaussian, $\sigma = 0.1$), and horizontal flips. Additionally, we generated 5,000 synthetic images per class using a Deep Convolutional GAN (DCGAN) with a carefully tuned set of hyperparameters to ensure realistic and diverse outputs. The generator consisted of five transposed convolutional layers with batch normalization and LeakyReLU activation, while the discriminator was structured as a deep convolutional network with four layers and spectral normalization to stabilize training, adding 15,000 synthetic images (5,000 each for 'knives,' 'pistols,' and 'guns'). The final dataset, after augmentations, reached approximately 39,000

images. For training, we split the dataset into 70% training (27,300 images), 20% validation (7,800 images), and 10% testing (3,900 images).

Fig. 2. Sample images from the OD weapon's dataset

2) GDXray Dataset: For X-ray classification, we use the publicly available GDXray dataset [10], which contains 8,000 images. Annotations in this dataset include bounding boxes and class labels (one for the presence of the threats and 0 for the absence of the threats). We added a custom Advanced X-Ray dataset class utilizing the albumentations library for it's adaptive image transformation capabilities. We applied various augmentations such as horizontal flipping and vertical flipping (50%) rotations upto ±15° (50%) probability, furthermore we apply brightness and contrast adjustments (30%) probability, with a range of ±0.2) to simulate fluctuations in X-ray beam strength and material density, coarse dropout (8 random 32 x 32 pixel regions dropped, (30%) probability) . We split the dataset into (70%) training (5,600 images), (20%) validation (1,600 images), and (10%) testing (800 images).

Fig. 3. sample images from the GDXray dataset

C. Model Pipeline

1) YOLOv11 Pipeline:

- **Preprocessing:** Images are resized to 640x640 and pixel values normalized to [0,1]. To improve model robustness, data augmentations are applied during training, including random horizontal flips, rotations ($\pm 15°$), and color jitter (brightness $\pm 20\%$, contrast $\pm 10\%$). These augmentations help the model generalize better to varied real-world CCTV footage.
- **Training:** The YOLOv11-m model, initialized with pre-trained COCO weights, is trained for 100 epochs with a batch size of 16. The learning rate starts at 0.001 and is dynamically adjusted using a cosine annealing scheduler, dropping to 0.0001 by the final epoch. The Adam optimizer is used with momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Mixed precision training is employed to optimize memory usage on the NVIDIA GTX 1650 GPU (4 GB VRAM), with training taking approximately 2000 hours. Early stopping is configured with a patience of 100 epochs, though it was not triggered due to consistent improvement.
- **Inference:** The model performs real-time detection on CCTV feeds at 30 frames per second (FPS). It achieves a mean Average Precision (mAP) of 0.96 at an Intersection over Union (IoU) threshold of 0.5, demonstrating strong accuracy in detecting objects like knives, pistols, and guns.

2) EfficientNetV2 Pipeline:

- **Preprocessing:** X-ray images are resized to varying dimensions (224x224, 300x300, and 384x384 across stages) and normalized using mean and standard deviation values typical of ImageNet (mean= [0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]). During training, augmentations such as horizontal/vertical flips, affine transformations, random brightness/contrast adjustments, coarse dropout, Gaussian noise, and blur are applied to enhance robustness.
- **Training:** An EfficientNetV2-M model, initialized with ImageNet pretrained weights, is fine-tuned over three progressive stages totaling 45 epochs (10, 15, and 20 epochs per stage). Batch sizes decrease from 64 to 32 to 16 as image resolution increases, optimized with the AdamP optimizer (initial learning rate 3e-4, decaying to 5e-5 across stages) and a cosine learning rate scheduler with warmup. The model leverages mixed precision training, EMA (Exponential Moving Average) weight averaging, and class-balanced sampling to handle imbalanced data, with training executed on a CUDA-enabled device - an NVIDIA GPU(GTX 1650), taking approximately 7 days.
- **Inference:** The trained model, using the EMA-averaged weights, classifies X-ray scans at an estimated rate of 15 ms per X-ray scan on the same NVIDIA GTX 1650 GPU.

D. Adaptive Frontend Integration

The Angular-based frontend integrates both systems:

- **CCTV Module:** Displays real-time video with bounding boxes around detected weapons, colored by class (red for guns, blue for knives, etc.). Users can adjust confidence thresholds (default 0.5) and zoom levels. Processing speed is 30 FPS on a low-range CPU (e.g., Intel i5-10th gen).
- **X-ray Module:** Shows classification results (threat/safe) with confidence scores. Users can review scans and generate reports. Processing speed is 2 seconds per scan.
- **Analytics:** Shows different types of Analytics with linecharts, piecharts and other diagrams for visualization and understanding.
- **Reporting:** Generates PDF/CSV logs with timestamps, threat types, and confidence levels.
- **System:** Displays various settings that can be changed and tailored for the use case of the organization.

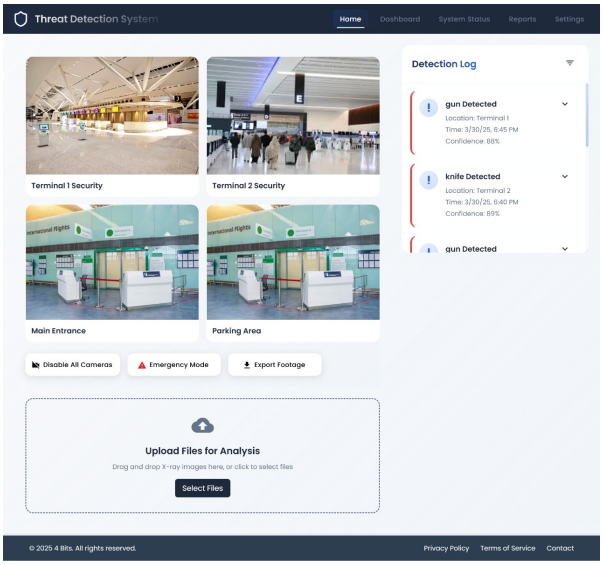

Fig. 4. Frontend Interface - Home

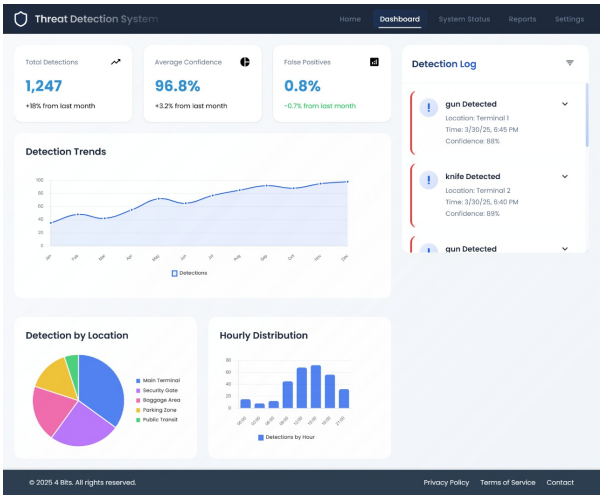

Fig. 5. Frontend Interface - Dashboard

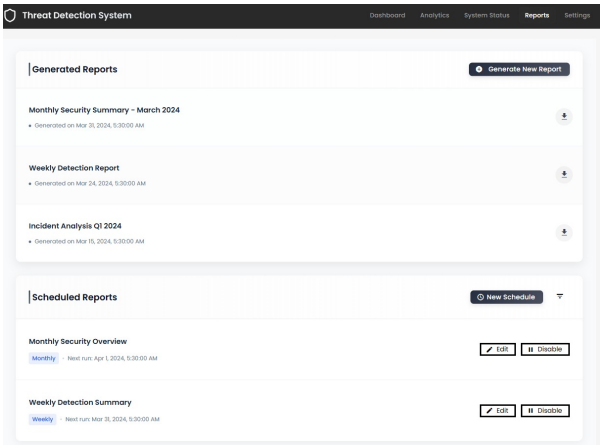

Fig. 6. Frontend Interface - Dashboard

IV. Result and Discussion

In this section, we present the results of the YOLOv11 model for real-time weapon detection and the EfficientNetV2 model for X-ray threat classifications. We have compared our models to the base paper's model against several metrics such as Accuracy, mAP (over various IoU) thresholds and F1 Score. We also included Visual representations of the detection findings and performance trends.

A. Performance Metrics for YOLOv11

The YOLOv11 model achieved an mAP of 0.960 at IoU 0.5, with class-specific mAPs: knives (0.960), pistols (0.990), and guns (0.920). At IoU 0.75, mAP was 0.880, and across IoU 0.5:0.95, it was 0.820. Precision was 0.935, recall 0.917, and F1-score 0.926. Ablation studies showed that including GAN-generated synthetic data improved mAP by 0.05 compared to training on the original dataset alone. Each video frame upto a total of 5 camera feeds can be actively monitored at a frame rate of 30 frames per second. The results of the evaluation and comparison with base papers are summarized in Table II.

TABLE II
Performance Comparison of YOLOv11 with Baseline Models

Metric	YOLOv8 [6]	SSD [7]	Our Model (YOLOv11)
mAP@0.5	0.780	0.543	0.960
mAP@0.75	0.680	–	0.880
mAP@0.95	0.720	–	0.820
Precision	0.739	0.852	0.935
Recall	1.00	–	0.917
F1 Score	0.857	–	0.926

B. Performance Metrics for EfficientNetV2

EfficientNetV2 achieved 99.44% accuracy, with an AUC of 0.999, precision of 0.999, recall of 0.995, and F1-score of 0.9934. The model correctly classified 99.0% of weapon scans and 99.4% of non-weapon scans. The false positive rate (FPR) was recorded at 0.238, indicating a strong ability to correctly classify non-weapons, while the false negative rate (FNR) remained as low as 1.65%, ensuring minimal weapon misclassification. Furthermore, with a mean inference time per image of just 15 ms, the model is well-suited for real-time applications, providing both speed and accuracy in high-risk surveillance environments. The results of the evaluation are summarized and compared to base papers in Table III.

TABLE III
Performance Comparison of EfficientNetV2 with HVGG19 Baseline

Metric	HVGG19 [16]	Our Model (EfficientNetV2)
Accuracy	95%	98.99%
AUC	0.950	0.999
Precision	0.72	0.999
Recall	0.739	0.995
F1 Score	0.857	0.9934

C. Training and Validation loss

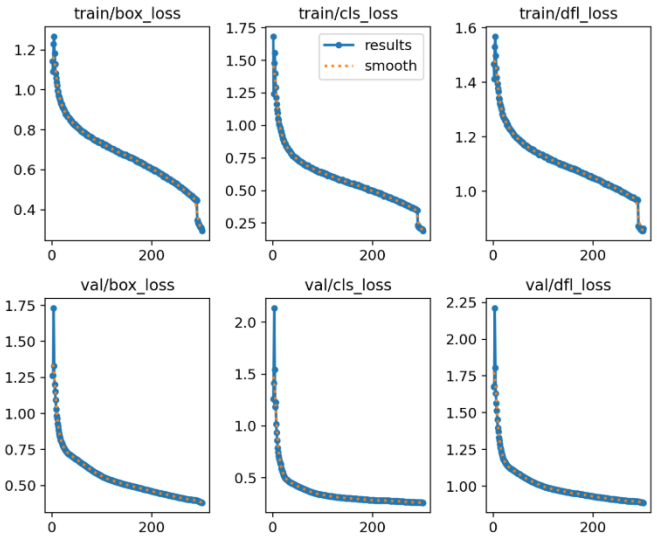

Fig. 7. YOLOv11 Training and Validation Loss

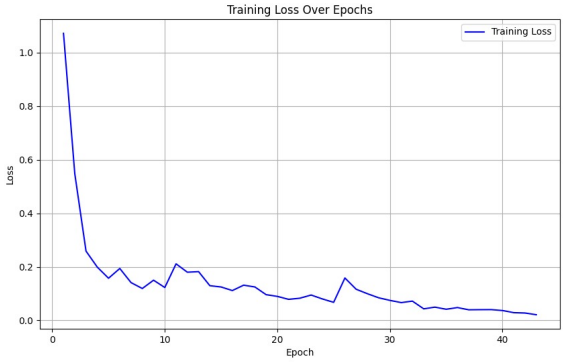

Fig. 8. EfficientNetV2 Training Loss

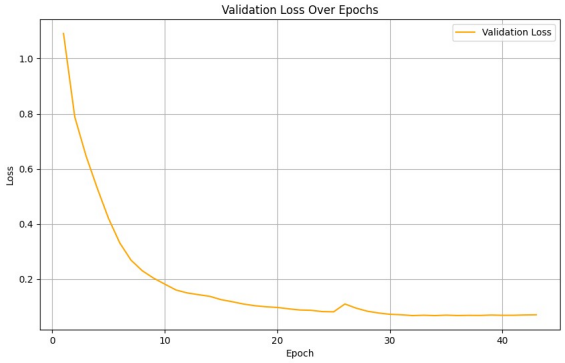

Fig. 9. EfficientNetV2 Validation Loss

D. Confusion Matrix

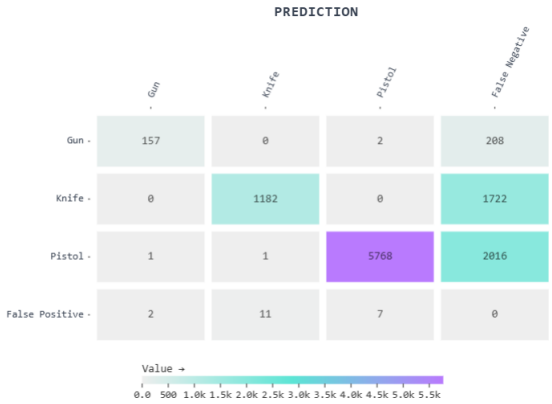

Fig. 10. YOLO v11 Confusion Matrix

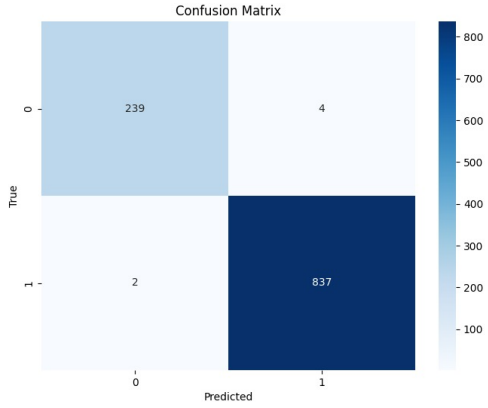

Fig. 11. EfficientNetV2 Confusion Matrix

E. Key Observation

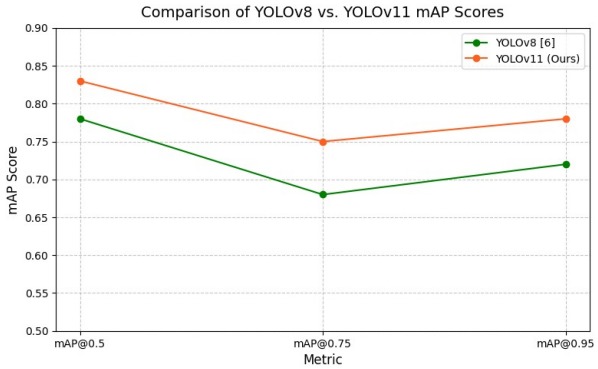

Fig. 12. YOLOv11 Key Observation

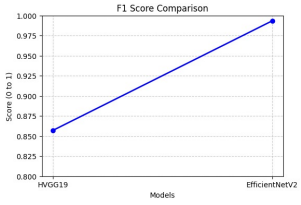

Fig. 13. EfficientNetV2 F1 Score Observation

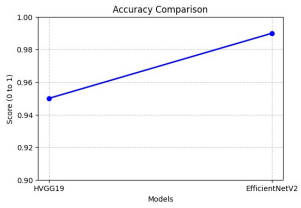

Fig. 14. EfficientNetV2 Accuracy Observation

V. Conclusion

In conclusion, integrating YOLOv11 for real-time weapon detection and EfficientNetV2 for real-time threat classification in X-ray scans establishes a complete, end-to-end threat detection system that addresses critical problems with current security systems. By utilizing YOLOv11's high-performance object detection, the system accurately detects weapons like knives, pistols, and guns. Furthermore, this system can also classify X-ray scans as threats or safe with good performance and still maintain being computationally light. The performance-efficient architecture of EfficientNetV2 makes this possible. To facilitate working with the framework, the development team adopted an Angular-based front-end application for the security personnel interface. The front end intended to present the model outputs by displaying bounding boxes drawn around detected threats. User-controllable zoom and confidence thresholding allow users to fine-tune analysis based on context. The system also incorporates a comprehensive reporting module that logs detailed audit trails, including time-stamped logs of detections, threat severity, and statistical summaries (e.g., threat frequency by category, and false positive rates). The reports can be sent in various formats (PDF, CSV) for application in compliance and operational audits. Deployment in high-stake environments like airports, government buildings, or public spaces reduces dependence on manual screening and human errors while increasing the speed of throughput without compromising safety. EfficientNetV2's power to execute resource-constrained hardware complements its ability to scale, ensuring integration with existing X-ray infrastructures to lower the barrier of adoption. Future research should reflect its limitations, including variations in training data and issues with applying results to novel classes of threats. Future work will aim to expand the dataset with more classes of weapons and incorporate multi-view 3D X-ray analysis and federated learning methods to enhance the robustness of the models across various scanning platforms

References

[1] Petrie et al., 2005 Petrie, C. V., Pepper, J. V., and Wellford, C. F. (2005). Firearms and violence: A critical review.

[2] Pittaro ML. School violence and social control theory: An evaluation of the Columbine massacre. Int J Crim Justice Sci. 2007;2(1):1-12.

[3] Kostinsky S, Bixler EO, Kettl PA. Threats of school violence in Pennsylvania after media coverage of the Columbine High School massacre: examining the role of imitation. Arch Pediatr Adolesc Med. 2001;155(9):994-1001

[4] Simon A. Application of fad theory to copycat crimes: quantitative data following the Columbine massacre. Psychol Rep. 2007;100(3):1233-1244.

[5] Fox JA. Review of All-American Massacre: The Tragic Role of American Culture and Society in Mass Shootings: By Eric Madfis & Adam Lankford, Eds., Philadelphia, PA: Temple University Press, 2023, 350 pages. Paperback, 34.95, 9781439923139; Hardcover: 115.50, 9781439923122; eBook: 34.95,9781439923146. 2024:1-3.

[6] Thakur, A., Shrivastav, A., Sharma, R., Kumar, T. and Puri, K., 2024. Real-Time Weapon Detection Using YOLOv8 for Enhanced Safety. arXiv preprint arXiv:2410.19862.

[7] S. Xu and K. Hung, "Development of an AI-based System for Automatic Detection and Recognition of Weapons in Surveillance Videos," 2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE), Malaysia, 2020, pp. 48-52, doi: 10.1109/ISCAIE47305.2020.9108816.

[8] Jegham, N., Koh, C.Y., Abdelatti, M. and Hendawi, A., 2024. Evaluating the evolution of yolo (you only look once) models: A comprehensive benchmark study of yolo11 and its predecessors. arXiv preprint arXiv:2411.00201.

[9] Abd El-Aziz, A. A., Mahmood A. Mahmood, and Sameh Abd El-Ghany. "A Robust EfficientNetV2-S Classifier for Predicting Acute Lymphoblastic Leukemia Based on Cross Validation." Symmetry 17.1 (2024): 24.

[10] Mery, D.; Riffo, V.; Zscherpel, U.; Mondragón, G.; Lillo, I.; Zuccar, I.; Lobel, H.; Carrasco, M. (2015): GDXray: The database of X-ray images for nondestructive testing. Journal of Nondestructive Evaluation, 34.4:1-12

[11] S. Tamboli, K. Jagadale, S. Mandavkar, N. Katkade and T. S. Ruprah, "A Comparative Analysis of Weapons Detection Using Various Deep Learning Techniques," 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2023, pp. 1141-1147, doi: 10.1109/ICOEI56765.2023.10125710.

[12] L. Jiao et al., "A Survey of Deep Learning-Based Object Detection," in IEEE Access, vol. 7, pp. 128837-128868, 2019, doi: 10.1109/ACCESS.2019.2939201.

[13] D. Liu, W. Wang, X. Wu and J. Yang, "EfficientNetv2 Model for Breast Cancer Histopathological Image Classification," 2022 3rd International Conference on Electronic Communication and Artificial Intelligence (IWECAI), Zhuhai, China, 2022, pp. 384-387, doi: 10.1109/IWECAI55315.2022.00081.

[14] Liang, K.J., Sigman, J.B., Spell, G.P., Strellis, D., Chang, W., Liu, F., Mehta, T. and Carin, L., 2019. Toward automatic threat recognition for airport X-ray baggage screening with deep convolutional object detection. arXiv preprint arXiv:1912.06329.

[15] M. Hayat, N. Ahmad, A. Nasir and Z. Ahmad Tariq, "Hybrid Deep Learning EfficientNetV2 and Vision Transformer (EffNetV2-ViT) Model for Breast Cancer Histopathological Image Classification," in IEEE Access, vol. 12, pp. 184119-184131, 2024, doi: 10.1109/ACCESS.2024.3503413.

[16] T. Morris, T. Chien and E. Goodman, "Convolutional Neural Networks for Automatic Threat Detection in Security X-Ray Images," 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 2018, pp. 285-292, doi: 10.1109/ICMLA.2018.00049.