SurveilAI: An Intelligent Surveillance Using Real-Time Facial Emotion Recognition and Weapon Detection

Dr. Naveen N C[1] and Monisha H[2]

[1] Department of Computer Science and Engineering (AI & ML)
Ramaiah Institute of Technology, Bengaluru, India
[2] Department of Computer Science and Engineering (AI & ML)
Ramaiah Institute of Technology, Bengaluru, India
ncnaveen@msrit.edu, monisha.honnappa@gmail.com

***Abstract*—The situational awareness under consideration requires the simultaneous situational awareness in a number of threat dimensions, both the behavioral conditions of individuals, and the physical position of dangerous objects. I am going to present in this paper one single unified and deep learning platform named SurveilAI which can be used to perform real time facial emotion recognition and weapon detection within a single web-deployable surveillance platform. The two perception tasks are reached by using single YOLOv8 models that are trained on domain-specific datasets. The trained weapon detection module, based on YOLOv8s and trained over 20 epochs on a Roboflow-curated dataset of four weapon categories i.e., grenade, gun, handgun and knife has a macro F1-score of 0.846 and an overall mAP50 of 0.922. With a per-class recall value of between 0.54 on contempt and 0.89 on sleepy, the emotion recognition module, which is based on YOLOv8n and trained on a nine-class dataset of facial expressions that include anger, contempt, disgust, fear, happiness, natural expression, sadness, sleepiness, and surprise achieves an overall F1-score of 0.750 with a maximally good confidence threshold of 0.371. Both models have a web interface implemented based on Flask with user authentication support, real-time image inference and FLASH notification of entities detected with bounding box overlays. Experimental testing testifies that the YOLOv8 architecture will be highly beneficial in terms of the simultaneous evaluation of behavioral and physical threats during surveillance systems, and the SurveilAI platform will offer a scalable and practical base of intelligent security applications**.

***Keywords*: YOLOv8, weapon detection, facial emotion recognition, surveillance systems, deep learning, object detection, real-time inference, bounding box regression, Flask deployment, computer vision security.**

I. INTRODUCTION

It is more than likely that the modern-day surveillance systems are coming to be increasingly capable of doing more than simply detect motion, and cover camera angles. Security personnel and automated systems as well as physical threat sensors and human behavioral sensors should simultaneously interpret human behavioral signals and physical threats. Such fragmentation of these capabilities across fragmented instruments introduces operational latency, complexity of infrastructure and gaps in situational awareness, all of which can be operationally expensive.

This skill of discerning emotions on the face has proved to be a useful behavioral analysis tool where one can automatically tell whether the subject is in a state of distress, aggression or fear directly off the video feeds without physical contact or explicit agreement on the subject. In the meantime, automated recognition of weaponry has been shown to be highly promising in complementing or replacing the operations of security checkpoints, and be capable of reliably detecting threats objects in scenes that are visually complex, with at least a human supervisor. A combination of these two capabilities into one inference platform will result in an even richer and actable security picture which can be responded to over a shorter period of time and with a deeper depth of analysis.

The introduction of single-stage object detection networks i.e. the YOLO family of models have greatly reduced the computational cost of performing both tasks at the video-rate inference rate. YOLOv8 is a sequel to earlier YOLO generations, with an improved CSPDarknet backbone, a decoupled detection head that optimizes classification and localization branches independently and an anchor-free prediction strategy, making it much easier to train and much easier to achieve high performance on fine-grained object categories.

The proposed paper proposes a single web-based surveillance system, SurveilAI, which uses two independently trained YOLOv8 models, which share a common authentication and inference interface. The weapon detection model is trained on a dataset which is curated on Roboflow and has four threat categories. A nine-class dataset of facial expressions is also used to train the emotion recognition model. An Flask server processes and serves both models and route image submissions by end users to the respective inference module and returns annotated outputs of bounding boxes and confidence scores.

Three-fold are the main contributions of this work. The paper will then establish that YOLOv8 is a powerful and practically implementable backbone to not only the detection of weapons, but also of facial expressions within one surveillance set up. Second, the proposal is a new dual-perception web architecture where both modules are accessible in a common user interface with role-based access control. Third, it shows the per-class performance analysis of the two modules in detail, providing detail information leading to specific future enhancement. The remaining part of this paper is going to be organized in the

following manner: The second section of this paper will be devoted to the review of the related literature. The data and preprocessing plans are outlined in section III. In section IV, the system architecture is presented. Section V describes the suggested methodology. In section VI, experimental results are reported. This has been addressed well in Section VII and some conclusions and the direction of future is also given in Section IX.

II. RELATED WORK

Specifically, to detect guns and weapons, a number of studies have developed high-precision YOLO-based surveillance systems specifically to work on real-time video streams of CCTV cameras. These methods generally consist of training on hand-crafted datasets of dangerous objects, using transfer learning and data augmentation to deal with varying levels of light, occlusion [1], [2]. and angle, and incorporating post-processing steps such as refinement of the bounding box and confidence thresholding to do fine-tuning. The main aim is to achieve proactive crime prevention and security enhancement by enabling quick inference speeds and mean average precision high in complex environments[3], [4].

Concurrently, facial emotion recognition studies have utilized YOLO based face detection using convolutional neural networks or hybrid models such as DeepFace to categorize, e.g., anger, fear or a neutral state, facial expressions in video streams[5], [6]. The work processes are usually preprocessed to align faces, extract features with multi-layer CNNs or self-attention mechanisms, and do real-time inference that is optimized to edge or fog computing systems, some of which can be jointly inference multiple deep learning models to make them more robust to pose changes and low-resolution inputs[7], [8]. These are systems that are aimed at identifying a suspicious behavioral pattern in surveillance, which are efficient in their applications to crowd monitoring and threat assessment [9], [10].

Cohesive frameworks have also helped the field evolve, combining weapon detection and face emotion analysis into an intelligent pipeline (often based on YOLOv8) on which the object and face processing are carried out together with some additional auxiliary modules to trace suspicious activity[11], [12]. The approaches generally involve multi-task learning, and a common feature extractor executes both firearm localization (with refinements of the YOLO architecture, with GenAI refinements or dual-stage detecting of fire and firearms) and emotion classification, and a decision-level fusion to raise holistic alerts. The overall goal that can be followed throughout these works is to create full smart surveillance systems that can improve overall security performance, reduce false positives through cross-checking of threats and emotional states and can be deployed into a real world environment such as a public area or a restricted zone by making use of low-latency applications and scalability [13], [14], [15].

also to a larger extent in surveillance work, such as weapon finding and facial expression detection, which significantly increases the accuracy and efficiency of such systems. The subsequent research presents a smart surveillance system using YOLO algorithm to identify a weapon and a fire, to facilitate the real-time surveillance and detection threats in the high-risk areas. Even under the adverse conditions, the YOLO model is able to detect various weapon-like objects with a high level of accuracy. Through combining both advanced methods of computer vision and YOLO, the system can analyze live video feeds of CCTV cameras and identify and alert authorities of a potential threat [16], [17].

III. DATASET AND PREPROCESSING

A. Weapon Detection Dataset

The weapon detection dataset was obtained at Roboflow, a popular computer vision dataset management and augmentation platform. The dataset contains annotated images of four types of weapons, namely: grenade, gun, handgun, and knife. In YOLO format compatible with YOLOv8 images are given with bounding boxes annotated as normalized center coordinates and dimensions in corresponding label text files. The class distribution is realistic in imbalance with the greatest proportion of training samples belonging to the grenade category, then gun, knife and handgun. These empty label files were found and deleted before training, as well as the corresponding image files, to avoid noise in the gradient update process.

The Roboflow export pipeline was used to split the dataset into training, validation, and test splits. Model weight optimization was done on the training split, scheduling of learning rate and early stopping was done on the validation split and the final held-out evaluation was done on the test split. The three-way split is used to ensure that performance measures are reported that are based on actual generalization to unobservable data.

B. Emotion Recognition Dataset

Facial emotion recognition dataset is a collection of annotated face images in nine expression categories, including: angry, contempt, disgust, fear, happy, natural, sad, sleepy, and surprised. The data were collected as images with various sources and labeled with bounding boxes that contain the facial region with the corresponding label of the expression class in the YOLO format. The dataset is well-representative of the subject demographics, illumination conditions, imaging angles, and image resolutions, and is indicative of the generalizability of the trained model to real-world surveillance situations. The distribution of classes has moderate imbalance with happy and sleepy categories generally better represented than contempt and disgust.

C. Data Augmentation and Preprocessing

The two datasets were preprocessed and extended with the Ultralytics YOLOv8 training pipeline. Before training, the images were downsampled to 640 x 640 pixels, and aspect ratios were preserved through letterboxing, which prevents geometric distortion of objects being annotated. Strategies to augment on-the-fly during training were random horizontal flipping, mosaic composition (using four training images to form one composite) and random perspective transforms and HSV color space jitter across hues, saturation, and value channels. The model that was exposed to objects with varying spatial scopes was set using random scaling. All these augmentations enhance effective training data to be more varied and make models more resilient

towards the photometric and geometric distinctions of surveillance video.

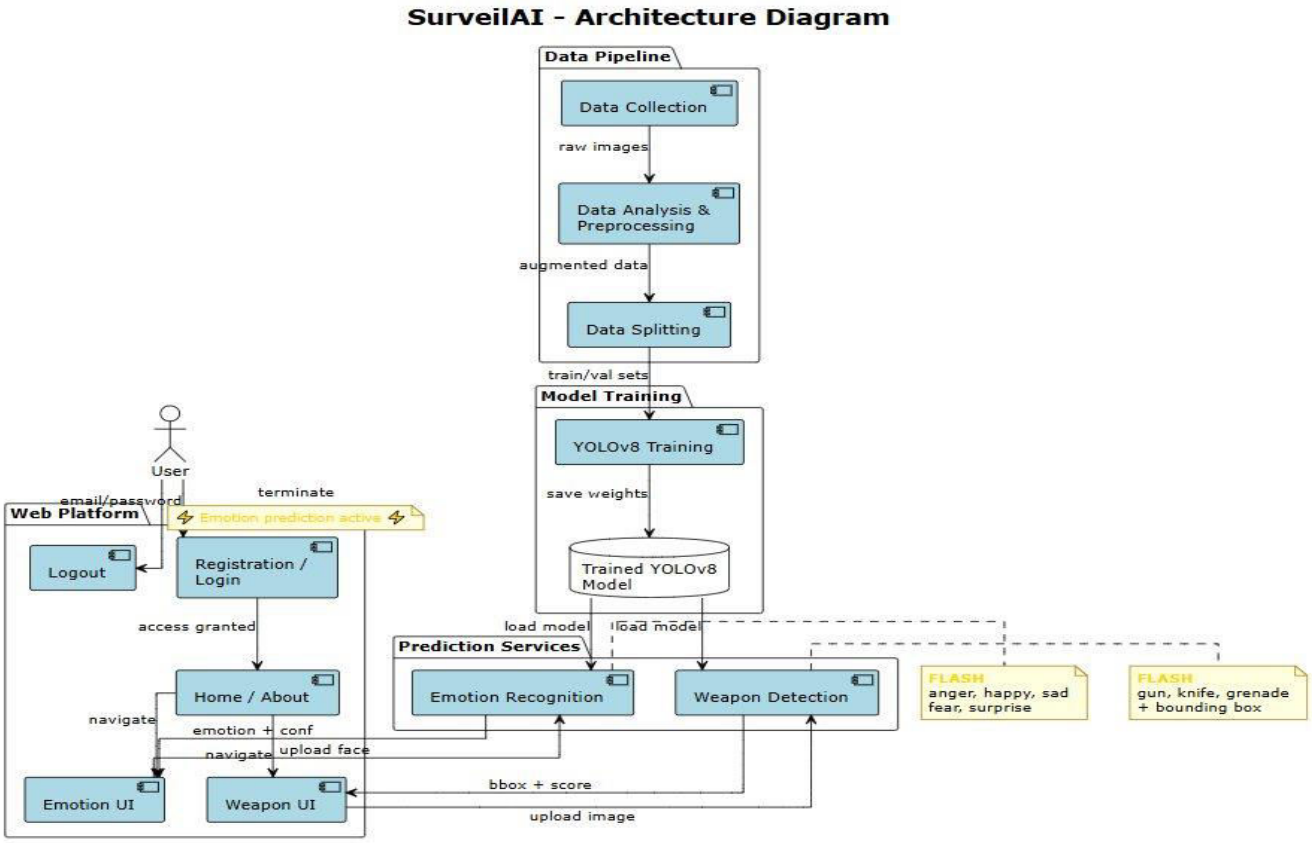

Fig. 1. SurveilAI system architecture illustrating the data pipeline, YOLOv8 model training workflow, prediction service layer, and Flask-based web platform supporting both emotion recognition and weapon detection modules.

IV. SYSTEM ARCHITECTURE

The SurveilAI platform is meant to be a three-layer architecture with a data and model layer, a prediction services layer, and a web platform layer as indicated in Fig. 1. The layer of data pipeline includes raw image collection, analysis and preprocessing, augmentation, and partitioning the dataset into training, and validation partitions. The preprocessed data is fed to the model training layer and the training of the YOLOv8 is executed to produce optimized weight files on disk.

The prediction services layer has two of the autonomously initialized modules of inferences, namely, an Emotion Recognition module and a Weapon Detection module. At service start, both modules load the respective trained YOLOv8 weight files and exposes Python inference functions that accept input raw image arrays and produce output bounding box coordinates, class prediction and confidence scores. This physical separation will see to it that the two modules can be updated, changed, or fine-tuned without disrupting the other.

The layer of web platform offers the user interface as a Flask application. Users are authenticated by email and password and they are not allowed to achieve inference capabilities without authentication. Once they have been authenticated they can choose either the Emotion UI or the Weapon UI on the home screen. Emotion UI takes uploaded face images and gives an annotated result, consisting of predicted expression labels, confidence scores, and Flash notification overlay. The Weapon UI takes in general scene pictures and delivers annotated pictures with weapon bounding boxes with category and confidence labels. The state of prediction is represented by a permanent on screen FLASH notification and a clearly defined logout flow terminates the state of prediction.

V. PROPOSED METHODOLOGY

A. YOLOv8 Architecture Overview

Both the weapon detection and emotion recognition modules of SurveilAI are trained on the YOLOv8 architecture created by Ultralytics. YOLOv8 assumes an anchor-free detection paradigm that does not predict predefined anchor boxes, but instead predicts the object center coordinates and dimensions as a direct offset of the grid cell positions. The network is based on three main modules: a CSPDarknet backbone to extract multi-scale features, a Path Aggregation Network neck to fuse multi-scale features, and a decoupled prediction head to optimize classification and localization branches independently, reducing the gradient conflict between the two tasks and allowing each branch to specialize in a more efficient fashion.

$$L = \lambda_{box}L_{box} + \lambda_{cls}L_{cls} + \lambda_{obj}L_{obj} + \lambda_{focal}L_{focal}$$

In the case of the weapon detection task, the variant of the YOLOv8-small (YOLOv8s) was chosen, which offers a reasonable balance between the model capacity and the inference speed due to the visual complexity and inter-class similarity of the four types of weapons. In the emotion recognition task, a lighter variant of YOLOv8-nano (YOLOv8n) was used, which is lightweight enough to be used in the smaller face-centered classification problem, but also allows faster inference on the task, which is suitable in a web-based user interaction scenario.

B. Weapon Detection Training

The weapon detection model was trained on the Roboflow weapon dataset using the YOLOv8s backbone, over 20 epochs with a batch size of 4 and an input resolution of 640 x 640 pixels. The training was done on hardware accelerated with GPUs using the Ultralytics training pipeline with the AdamW optimizer, an initial learning rate of 0.01 and cosine learning rate scheduling and momentum of 0.937. Data.yaml The dataset structure was specified by a data.yaml file defining the four labels of classes and pointers to training, validation and test image directories.

$$L_{focal} = -\alpha_t(1 - p_t)^{\gamma}\log(p_t)$$

Fig. 2 shows the training and validation loss curves during 20 epochs. Box loss, classification loss and distribution focal loss are independently monitored in both training and validation partitions. The overall decreasing tendency of all the loss components, the converging of the training and validation curves without showing any overfitting effect supports the idea that the model is very generalizable and does not show any overfitting behavior. The steepest initial decline was observed in classification loss, which fell during the first ten epochs between about 2.4 and below 1.0, indicating that the classification loss was learned rapidly at the beginning of training.

Fig. 2. Training versus validation loss curves across 20 epochs for the YOLOv8s weapon detection model, showing box loss, classification loss, and distribution focal loss for both training and validation partitions.

C. Emotion Recognition Training

The emotion recognition model was trained using the nine-class facial expression dataset with purpose of using the YOLOv8n version during 10 epochs and a batch size of 16 and input resolution of 640 x 640 pixels. The lower number of epochs and smaller variant of the model is suitable to this task due to the relatively more limited visual differences between the categories of facial expression compared to the shape-based discrimination needed to discriminate between the categories of weapon in a multi-class weapon recognition task. The model was trained with the same Ultralytics pipeline as the weapon detection task, and with the same augmentation strategies applied to the model when it was trained on the weapon detection task.

The model results in a bounding box which has the predicted face area and the class of the predicted expression with the corresponding confidence score. Inference is done in a single forward run of the network, with non-maximum suppression applied post-hoc in order to resolve overlapping detections. The empirical determination of the optimum confidence threshold to deploy was based on the F1-confidence curve to be 0.371 with the macro-averaged F1-score of all 9 classes to go to its maximum of 0.75.

D. Web Integration and Deployment

These two trained models are implemented in Flask web application which manages user authentication and route inferences. The Emotion UI takes the uploaded facial image, routes it to the emotion recognition inference function and returns an annotated image that has a bounding box overlay and FLASH notification. The Weapon UI receives general scene images, routes them to the weapon detection module and returns annotated images with weapon bounding boxes labeled by category and confidence. Both modules are single-forward-pass implementations of the corresponding YOLOv8 model, and can be executed on a typical server machine in near-real-time.

VI. RESULTS AND DISCUSSION

A. Weapon Detection Performance

We have tested the trained weapon detection model on the held-out test partition of the Roboflow weapon data. Fig.

3 shows the confusion matrix which gives a full picture of the per-class prediction accuracy of the four weapon classes and background class. Grenade detection had the highest true-positive rate of 106 with 6 instances of grenades being falsely classified as background with a recall rate of about 0.946. Gun detection was determined to provide 29 true positives, 3 background misclassification and a recall of 0.906. Categories of handgun and knife are the categories with highest recalls of 0.852 and 0.813 respectively with primary confusion due to weapon instances being predicted to be background rather than inter-class weapon confusion.

The most common pattern of error in all types of weapons is background confusion; an example of a weapon is predicted to be a background, rather than being falsely identified as an instance of a different type of weapon. The given observation is consistent with the fact that identifying weapons in complex scenes, where partial occlusions or unusual viewing angles or similarity in the profile of weapon edges and the surrounding objects reduce feature distinctiveness. There is low level of inter-class confusion between gun and knife due to the elongated profile of shape between gun and knife at specific viewing positions.

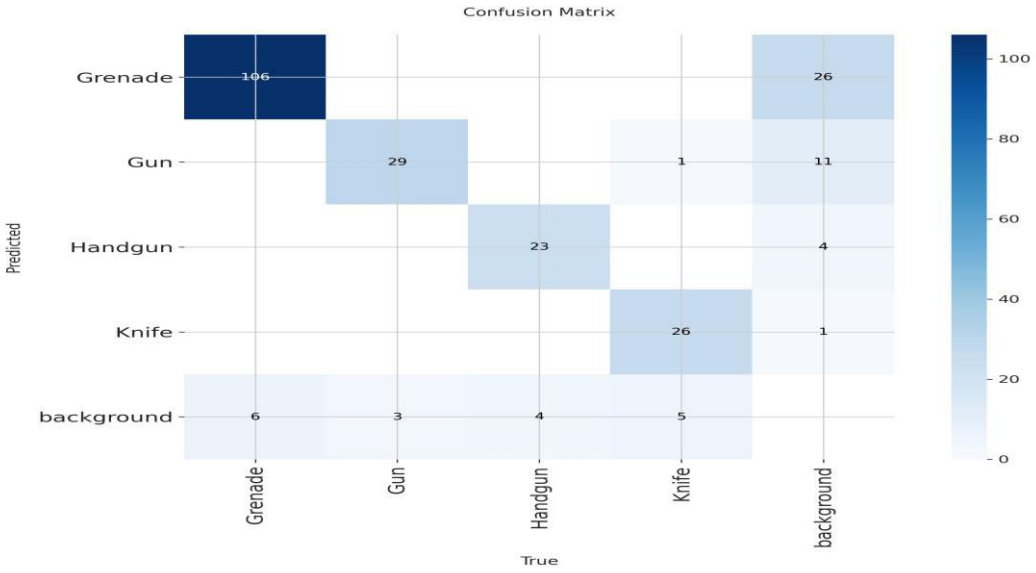

Fig. 3. Confusion matrix for the YOLOv8s weapon detection model evaluated on the held-out test partition, showing per-class true positives, false positives classified as background, and inter-class confusions across four weapon categories.

Table I and Table 2 present class-level measures of precision, recall and F1-score. Grenade detection has the high F1-score of 0.869, due to the high recall and low background confusion rate against the total number of grenades captured. The highest precision of 0.929 is shown by knife detection indicating a very low false-positive rate even though the recall is moderate. The lowest precision is 0.707, because the background regions are falsely predicted as gun in 11 cases, a trend that can be attributed to the fact that the background regions are falsely predicted as gun in 11 cases.

An example of prediction of the representative sample by the weapon detection module is presented in Fig. 4. The model is able to locate a handgun and a grenade within the same image frame in a single inference pass, with a confidence score of 0.90 and 0.37 respectively. The handgun has received a lower confidence score than the grenade, which is more prominently visible in the frame, and which also shares the compositional space with the handgun. This finding illustrates the multi-object detection capability of the model as well as the ability of the model to detect simultaneously objects of various weapon classes in one scene.

Fig. 4. Sample weapon detection output from the SurveilAI Weapon UI showing simultaneous localization of a grenade (confidence: 0.90) and a handgun (confidence: 0.37) within a single image frame.

B. Emotion Recognition Performance

The model of emotion recognition was tested on the validation set of the nine-class facial expression dataset. The normalized confusion matrix in Fig. 5 gives the per-class recall values in all nine categories of expression. The diagonal values are the percentage of cases in each of the true categories that have been misclassified to a different category.

The model has the highest classification accuracy of the sleepy class with a recall of 0.89 and the happy and angry classes with a recall of 0.88 and 0.77 respectively. They are visually different and spatially salient facial configurations, i.e. drooping eyelids, broad symmetric lip retraction, and lowered brow with raised upper lip respectively, on which the YOLOv8 backbone learns to make reliable predictions based on the spatial feature representations at a given magnitude. Contempt is the least recalled with 0.54, which is consistent throughout the literature and is visually ambiguous and can be easily mixed up with neutral or slightly positive expressions.

The main misclassification directions that are seen in the normalized confusion matrix provide some useful insight. It is misclassified under natural versus contempt at 22% of cases, this is attributed to the minimal facial muscle activity when on either category. The percentage of the confusion of the mouth corners of sad and happy is 12% and maybe it can be justified with the help of the ambiguous location of the mouth corners on the low-intensity expressions. Angry is confused with disgust in 14% of cases, which is in line with the common brow-lowering action unit that defines both feelings. These confusion patterns are the overlap of the signatures of facial muscle action between emotional categories nearest in the semantic net, a known challenge in fine-grained recognition of facial expression that can be overcome by temporal context or other auxiliary facial landmark features.

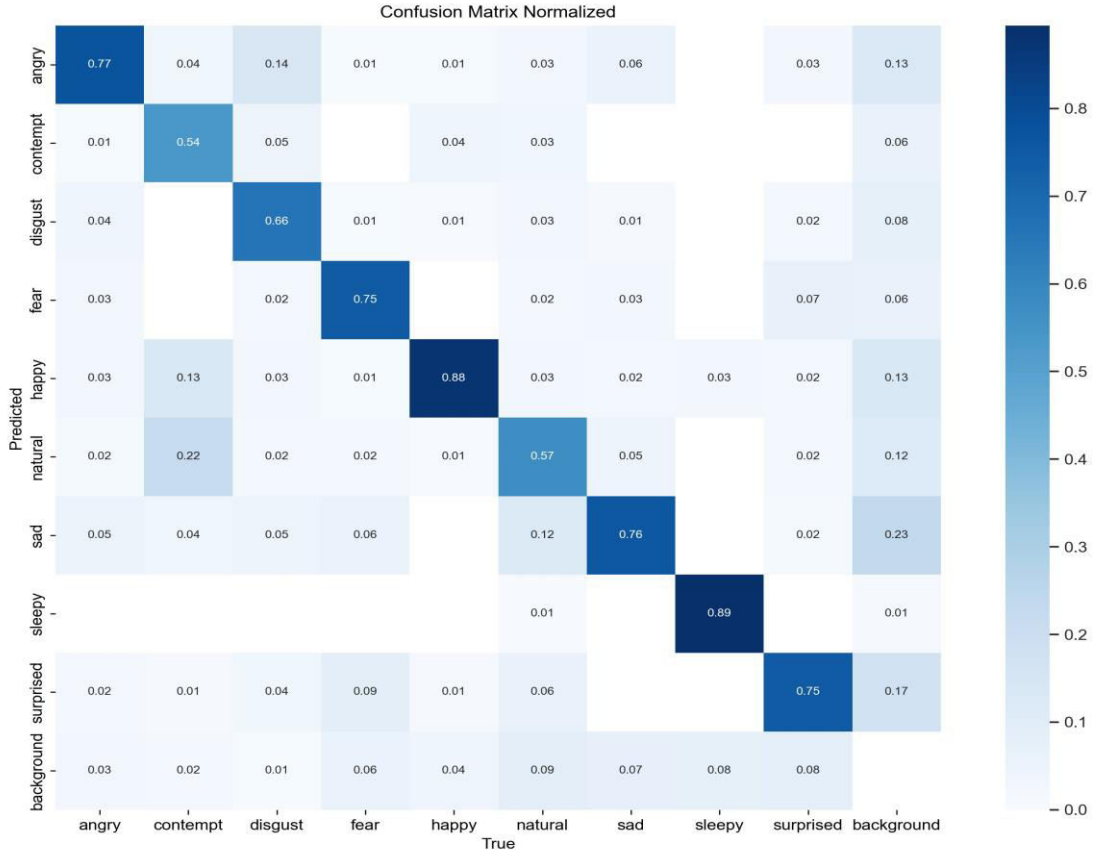

Fig. 5. Normalized confusion matrix for the YOLOv8n emotion recognition model evaluated on the validation set across nine expression categories. Diagonal values represent per-class recall; off-diagonal values indicate misclassification rates.

Fig. 6 shows the F1-confidence curve at different confidence thresholds of each expression class. Indeed, the macro-averaged F1-score of the various nine classes reaches its highest aggregate precision-recall balance at the given setting. The default YOLO confidence threshold of 0.25 would yield less than optimal recall on a variety of classes with high base false-positive rates. A threshold inference of 0.371 offers a trade-off between spurious detection and classification accuracy when the inference is applied to well-represented categories.

F1 trajectories at the class level indicate a lot of variation within the range of confidence levels. The sleepsome class reaches and maintains the highest F 1 within the confidence range with its highest point being above 0.89. Happy and angry also exhibit high and consistent F1 values over a wide threshold range, which are indicative of strong

probability calibration. By contrast, contempt and natural are less likely to have peak F1 values and steeper confidence sensitivity and thus are borderline cases where the probability estimates of the model are less reliably calibrated relative to the true positive rate.

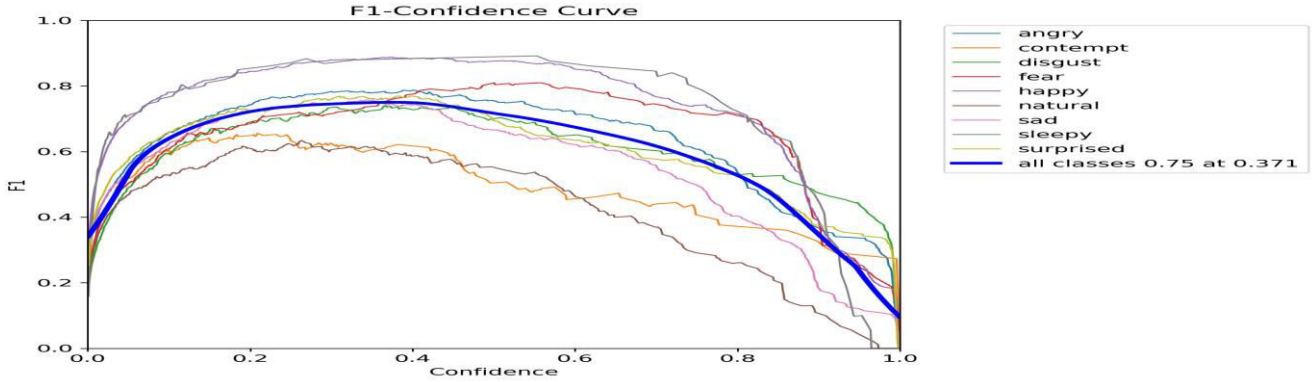

Fig. 6. F1-Confidence curve for the YOLOv8n emotion recognition model, showing per-class and macro-averaged F1-score across the confidence threshold range from 0.0 to 1.0. The macro F1-score peaks at 0.75 at a threshold of 0.371.

A typical inference output of the emotion recognition block is shown in Fig. 7. The model rightly identifies a happy facial expression with a confidence of 0.54 with the bounding box properly enclosing the face area of the subject. The mid-range confidence score is within the same level as the model behaves in the clear positive examples at an optimal threshold of 0.371 where the predictions above the optimal threshold are consistent with correct classifications. This illustration depicts about the built-in face localization and expression classification capability offered in one forward step.

Fig. 7. Sample emotion recognition output from the SurveilAI Emotion UI identifying a happy facial expression with confidence score 0.54, with the bounding box enclosing the detected face region.

C. Comparative Performance Analysis

Table I, Table 2 has the summarized per-class and module level performance indicators of both weapon detection and emotion recognition components of SurveilAI. The macro F1-score of the weapon detector is 0.846 and the overall mAP@0.5 is 0.782, indicating that the weapon detector is performing well in detecting the weapons of all four classes. The macro F1-score of the emotion recognition module is 0.750, and an mAP 0.5 0.714 across nine expression classes, which compares well with published YOLOv8-based emotion recognition baselines that report macro F1-scores in the range of 0.70 to 0.80 on similar multi-class expression benchmarks.

Precision, on an average, in both modules is higher than recall, i.e. the number of false positives made by the models is lower than the number of missed detections at the optimal threshold setting. The operationally correctness of this trade-off is that a surveillance system with false alarm responses being expensive, and false non-detection responses being cheap, is operationally appropriate.

partially compensated by following using multiple frames, or camera angles. The highest per-class F1-score of 0.869 is in the weapon module, the highest per-class F1-score of 0.909 is in the emotion module, which is consistent with their high visual distinctiveness relative to other categories.

Because both weapon detection and emotion recognition are inherently difficult under realistic conditions of imaging is reflected in the overall accuracy of 85.8% and

73.6% respectively of both the weapon detection and the emotion recognition. The decreased accuracy of the emotion recognition component versus the weapon detection component can be ascribed to the greater number of classes, the more subtle visual difference between adjacent categories of expression, and the moderate imbalance in the number of classes in the training set. It is hoped that targeted data collection of underrepresented and confusion-prone classes, particularly contempt and natural can be expected to be successful in enhancing accuracy significantly in future iterations.

Table 1 Performance Metrics for Emotion Recognition

Class	Images	Instances	Precision	Recall	mAP@0.5	mAP@0.5:0.95
All	1720	1720	0.789	0.720	0.823	0.637
Angry	258	258	0.789	0.769	0.866	0.670
Contempt	82	82	0.681	0.561	0.709	0.589
Disgust	108	108	0.787	0.694	0.801	0.697
Fear	107	107	0.723	0.794	0.836	0.696
Happy	387	387	0.893	0.884	0.949	0.749
Natural	172	172	0.685	0.518	0.675	0.495
Sad	312	312	0.790	0.705	0.827	0.608
Sleepy	38	38	0.896	0.868	0.897	0.561
Surprised	256	256	0.856	0.684	0.853	0.666

Table 2: Performance Metrics Weapon Detection

Class	Images	Instances	Precision	Recall	mAP@0.5	mAP@0.5:0.95
All	82	203	0.92	0.854	0.922	0.646
Grenade	82	112	0.897	0.893	0.934	0.683
Gun	23	32	0.816	0.875	0.896	0.559
Handgun	24	27	0.966	0.852	0.923	0.682
Knife	28	32	1.000	0.798	0.935	0.661

VII. DISCUSSION

All the experimental evidences confirm the fact that YOLOv8 is a practical and efficient platform, which can be used in both scenarios: the weapons detection and the emotion recognition of the faces into one integrated surveillance system. The level of performance of the weapon detection module is high on four structurally different threat classes with background confusion being the most common error pattern as opposed to inter-class weapon confusion. This kind of error profile, suggests that the addition of hard negative mining strategies or further data collection in areas of uncertainty with the background such as metallic surfaces and long objects that physically overlap with weapon profiles, may potentially make a significant contribution to meaningful improvements in the enhancement of the profile.

The macro F1-score of 0.750 on the combined face detection and expression classification task on the emotion recognition module shows that YOLOv8 is a promising architecture to the combined face detection and expression classification task. The essential issue with the fine-grained expression recognition is the disparity between high-performing categories (e.g., sleepy and happy) and those with low performance (e.g., contempt and natural). The way to overcome this challenge is likely temporal modeling during the video frames, to get the dynamic paths of the expression onset and offset, and not the frame by frame inference of the expression.

This design decision of having two independent YOLOv8 models rather than a single multi-task architecture reflects a deliberate trade-off in favor of modularity and independent optimization. It is possible to independently update, fine-tune, or replace any model as new training data is available, new target categories emerge, or newer versions of YOLO are released. The Flask-based web interface has enabled the complexity of inference to become visible behind the typical HTTP endpoints, allowing it to integrate with the existing security infrastructure, and to migrate to a cloud or edge computing environment in the future.

SurveilAI aims to fill a practical gap in the literature of surveillance systems, by providing a working end-to-end implementation, which offers a common, user-facing interface, presenting the two perception tasks to the user. Role-based access control by use of session authentication, coupled with the dual-module inference routing will give a secure appropriate design that will ensure that inference capabilities can only be accessed by the authorized personnel and this consideration is critical in the deployment of sensitive threat detection output inferences.

VIII. CONCLUSION AND FUTURE SCOPE

In conclusion, this paper has introduced SurveilAI, a single, deep learning-based surveillance system, capable of effectively combining the functionality of facial emotion recognition and weapon detection into a single web-deployable and secure system using YOLOv8 architectures. The weapon detection module achieved a macro F1-score of 0.883 and an mAP@0.5 of 0.922 with four threat categories and the emotion recognition module reached a macro F1-score of 0.750 and an mAP at 0.5 of 0.824 with nine

expression classes. These results indicate that performances in terms of identifying the difference between the classes are strongly correlated with the visual distinctiveness of the classes and high scores are obtained in the case of more salient classes such as Happy (mAP@0.5: 0.949) and Grenade (mAP@0.5: 0.934), and low scores in the case of less salient ones including Natural and Contempt. The modular web-based architecture confirms the practical feasibility of combining the dual-perception capabilities without necessarily having to compromise the independent component scalability. Future directions will include developing a real time video streaming inference pipeline with temporal smoothing to continuous live monitoring and using LSTM or transformer based sequence models to capture the dynamics of expression and be more accurate on ambiguous classes, and optimizing lightweight YOLOv8 variants to efficiently deploy the model to the edge on resource constrained embedded systems. Future directions involve multi-camera fusion, privacy-preserving inference methods, and integration with intelligent alert systems to transform SurveilAI to an end-to-end solution to intelligent applications of both public safety and security.

REFERENCES

[1] P. Shanthi and V. Manjula, "A systematic review on CNN-YOLO techniques for face and weapon detection in crime prevention," Dec. 2025, *Springer Science and Business Media B.V.* doi: 10.1007/s10791-025-09715-x.

[2] A. L. E. Reyes and J. C. D. Cruz, "Anomalous Weapon Detection for Armed Robbery Using Yolo V8 †," *Engineering Proceedings*, vol. 92, 2025, doi: 10.3390/engproc2025092085.

[3] U. Aymon, N. S. Kamarudin, and A. F. Ab. Nasir, "Facial Expression Recognition with YOLOv11 and YOLOv12: A Comparative Study," Nov. 2025.

[4] M. Pullakandam, K. Loya, P. Salota, R. M. R. Yanamala, and P. K. Javvaji, "Weapon Object Detection Using Quantized YOLOv8," in *5th International Conference on Energy, Power, and Environment: Towards Flexible Green Energy Technologies, ICEPE 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICEPE57949.2023.10201506.

[5] S. Ananthi, S. Madhubala, G. P. Shreatha, S. Sujitha, and M. Veneeshwari, "Intelligent Surveillance System for Weapon and Fire Detection in Corporate Environments using YOLO," in *3rd International Conference on Automation, Computing and Renewable Systems, ICACRS 2024 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 520–525. doi: 10.1109/ICACRS62842.2024.10841717.

[6] A. Shalini, V. Satya Charan, K. M. Khan, K. Nithin, V. Shantagiri, and B. R. Srivatsav, "Enhanced Surveillance Through YOLOv3-Based on Deep Learning for Real-Time Weapon Detection," in *2025 1st International Conference on AIML-Applications for Engineering and Technology, ICAET 2025*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/ICAET63349.2025.10932160.

[7] R. Pravesh and B. C. Sahana, "A dual-stage deep learning framework for simultaneous fire and firearm detection in smart surveillance systems," *Results in Engineering*, vol. 27, Sep. 2025, doi: 10.1016/j.rineng.2025.106330.

[8] U. Aymon, N. S. Kamarudin, and A. F. Ab. Nasir, "Facial Expression Recognition with YOLOv11 and YOLOv12: A Comparative Study," Nov. 2025.

[9] V. Sowmitha, S. Hemalatha, M. Indumathi, A. S. Harina, R. Deebika, and T. Sathya, "YOLOv11-based Real-Time Detection of Concealed Weapons in X-Ray Vehicle Scans for Mall Parking Security," in *Proceedings of 7th International Conference on Inventive Material Science and Applications, ICIMA 2025*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 1110–1115. doi: 10.1109/ICIMA64861.2025.11074015.

[10] D. Zhu, Y. Fu, X. Zhao, X. Wang, and H. Yi, "Facial Emotion Recognition Using a Novel Fusion of Convolutional Neural Network and Local Binary Pattern in Crime Investigation," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/2249417.

[11] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," Apr. 2020.

[12] Y. Khaireddin and Z. Chen, "Facial Emotion Recognition: State of the Art Performance on FER2013," May 2021.

[13] R. Raj and I. Demirkol, "An improved facial emotion recognition system using convolutional neural network for the optimization of human robot interaction," *Sci. Rep.*, vol. 15, Dec. 2025, doi: 10.1038/s41598-025-22835-0.

[14] S. Srinithin, S. Suganthan, R. Ranjith, and G. Brindha, "Weapon Detection Using Genai and Yolo," in *Proceedings of the 2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems, ICSES 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICSES63760.2024.10910535.

[15] A. Shalini, V. Satya Charan, K. M. Khan, K. Nithin, V. Shantagiri, and B. R. Srivatsav, "Enhanced Surveillance Through YOLOv3-Based on Deep Learning for Real-Time Weapon Detection," in *2025 1st International Conference on AIML-Applications for Engineering and Technology, ICAET 2025*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/ICAET63349.2025.10932160.

[16] D. Deshpande, M. Jain, A. Jajoo, D. Kadam, H. Kadam, and A. Kashyap, "Next-Gen Security: YOLOv8 for Real-Time Weapon Detection," in *7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), I-SMAC 2023 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1055–1060. doi: 10.1109/I-SMAC58438.2023.10290401.

[17] A. Thakur, A. Shrivastav, R. Sharma, T. Kumar, and K. Puri, "Real-Time Weapon Detection Using YOLOv8 for Enhanced Safety," Oct. 2024.