Grenze International Journal of Engineering and Technology, June Issue

An Adaptive Pooling YOLO-based Multimodal Detection Algorithm for Secure Automated Interview Assessment

Princy S[1], Prithika M[2], Randhini E L[3] and Lavanya V[4]
[1-4]Kumaraguru College of Technology/Computer Science Engineering, Saravanampatti, Coimbatore, India
Email: princysesu40@gmail.com, prithikha.22cs@kct.ac.in, randhini.22cs@kct.ac.in, lavanya.v.cse@kct.ac.in

***Abstract*— The current recruitment system requires efficient remote technical interview processes that provide secure and scalable solutions while reducing bias. Traditional interview methods are time-consuming and often produce inconsistent results due to their reliance on human judgment. Online recruitment has also introduced challenges such as candidate authentication, cheating detection, and effective evaluation of technical skills. Advancements in Artificial Intelligence (AI) and Deep Learning (DL) have enabled automated assessment systems with improved decision-making, contextual understanding, and real-time interaction capabilities. This research proposes an automated interview framework that integrates Natural Language Processing (NLP), computer vision, and speech technologies to conduct secure technical evaluations based on job requirements. The system dynamically generates interview questions using GPT-based language understanding and monitors candidates through an improved AP-YOLO model for real-time monitoring and cheating detection. A CNN model trained on the FER2013 dataset is used for facial emotion recognition to support effective evaluation while maintaining candidate confidence. The framework enables natural interaction through speech-to-text and text-to-speech technologies while ensuring security through browser activity monitoring, multiple-person detection, and visual object recognition. The proposed system improves recruitment efficiency by automating scoring, generating adaptive questions, and providing unbiased evaluation without human intervention, while future work will focus on semantic answer evaluation, multilingual support, and advanced behavioral analysis to enhance AI-based decision making.**

***Index Terms*— Artificial Intelligence, Automated Interview System, Deep Learning, AP-YOLO Algorithm, Natural Language Processing, Emotion Recognition, FER2013 Dataset, Cheating Detection, Speech Recognition, Technical Recruitment.**

I. INTRODUCTION

The job recruitment process has experienced a complete transformation because new digital methods emerged together with remote workforces; companies now use virtual hiring systems to conduct rapid and low-cost screenings of large applicant pools. Organizations that adopt virtual hiring systems still rely on human interview systems which operate with inconsistent candidate assessment standards and different interview scheduling practices and interviewer bias tendencies, causing organizations to spend excessive time on their hiring processes.

Grenze ID: 01.GIJET.12.2.198_24

The technical recruitment field requires pre-screening of technical skills and problem-solving abilities together with behavioral characteristics because these skills are essential for success in the recruitment process. The introduction of Artificial Intelligence (AI) and Deep Learning (DL) technologies enables recruitment teams to automate and enhance specific recruitment tasks which occur during the initial phases of the hiring process. New AI systems can analyze extensive data sets to achieve automatic system adjustment based on individual candidate responses while delivering an impartial scoring system as a standard assessment tool. Natural Language Processing (NLP) technology allows intelligent systems to process candidate natural language input while creating interview questions that match specific job requirements. Automated interview systems can conduct natural dialogues with candidates through speech recognition and synthesis technologies which enable the system to understand and generate human language. The main problem which affects remote work environments is assessment integrity. Online interviews face greater risks of cheating because participants can cheat by using forbidden resources or pretending to be other people or receiving help from outside sources when they should not. The automated monitoring systems which exist today lack sufficient capabilities to detect and respond to all forms of malicious activities which may occur. The use of vision-based remote monitoring techniques has increased in employment-related computer applications because these technologies provide effective remote observation solutions. The candidate's environment can be observed in real time through advanced object detection algorithms which enable tracking of suspicious objects and identification of multiple individuals and detection of irregular activity patterns. Candidate psychological evaluation now requires assessment of behavioral and emotional aspects along with traditional technical skill measurement methods. The recruiter uses emotional intelligence values and confidence scores and stress indicators to determine which job roles suit the candidate best. Deep learning systems perform facial emotion recognition to enable automatic systems to identify facial expressions and determine emotional states throughout the interview process. The use of standardized emotion recognition datasets has enabled researchers to develop this capability. The existence of standardized datasets for emotion recognition has led to the development of trustworthy behavior analysis models which enhance accuracy in interview assessments. It is in this context that we have proposed to develop an intelligent AI system for automatic interview which will utilise various technologies addressing the challenges mentioned above. The framework includes the following four modules: (1) Adaptive question generation with Natural Language Processing, (2) interactive communication based on speech technologies, (3) real-time integrity monitoring informed by advanced object detection algorithms and (4) emotion analysis through deep learning models fed from facial expression datasets. The platform also includes built-in auto-response grading, browser-activity monitoring, automatic scoring templates and dynamic reporting features to facilitate analysis of test performance. The research aims to create an automated interview system which operates at a large scale while maintaining security and unbiased assessment capabilities to improve recruitment processes and ensure equal treatment of all candidates. The presented system uses AI to minimize human work while implementing advanced monitoring and evaluation tools to tackle the major challenges that remote hiring operations face in current times. Our intelligent service for technical recruitment and hiring decisions establishes a fair and transparent and efficient system which advances Human Resource Technology development in the digital domain.

II. LITERATURE SURVEY

The growing demand for scalable and efficient recruitment has led to the adoption of AI-based automated interview systems. Traditional interview methods rely on manual evaluation, resulting in subjectivity, high cost, and inefficiency in large-scale hiring. AI and Machine Learning (ML) techniques enable automated candidate assessment, standardized evaluation, and improved hiring efficiency [1], [2]. Natural Language Processing (NLP) plays a key role in automated interviews by generating adaptive questions and evaluating candidate responses. Previous works have demonstrated the use of NLP for resume analysis and intelligent question generation, improving the accuracy of skill assessment [3], [4]. Speech recognition technologies are widely used to convert spoken responses into text and enable natural interaction between candidates and automated systems [5], [6]. To ensure interview integrity, computer vision-based proctoring systems have been developed to detect suspicious activities such as multiple persons and unauthorized object usage during online interviews [7], [8]. Recent advancements in deep learning have improved object detection performance, particularly with YOLO-based models, which are suitable for real-time monitoring applications [9]. Enhanced models such as AP-YOLO provide better detection of small and occluded objects. In addition, facial emotion recognition using CNN models and datasets like FER2013 enables behavioral analysis by identifying candidate emotions such as stress and confidence [10], [11]. Despite these advancements, existing systems often lack integration, transparency, and robustness. Many solutions focus only on assessment or monitoring, rather than combining both. This work addresses these limitations by proposing a unified AI-based automated interview system integrating NLP, speech processing, object detection, and emotion analysis for secure and unbiased evaluation.

III. Overview of the Automated Interview and Recruitment Challenge

The rapid growth of digital hiring and remote work has made technical recruitment a complex and resource-intensive process. Organizations receive a large number of applications, making traditional interview methods inefficient, time-consuming, and difficult to scale. Manual evaluation introduces inconsistencies, scheduling challenges, and interviewer bias, affecting the fairness and quality of hiring decisions. A major challenge in recruitment is the lack of standardized evaluation procedures. Different interviewers may assess candidates differently, leading to inconsistent results and potential selection errors. Additionally, remote interviews introduce risks such as cheating, impersonation, and the use of unauthorized resources, which existing systems fail to detect effectively. Another key challenge is the need to evaluate both technical skills and behavioral traits. Traditional methods rely on human observation, which may not provide a complete or objective assessment of candidate performance. Factors such as confidence, engagement, and emotional state are difficult to measure accurately using manual approaches. Artificial Intelligence (AI) and Deep Learning (DL) provide solutions to these challenges by enabling automated, scalable, and unbiased evaluation. NLP supports intelligent response analysis, while speech recognition enables seamless interaction. Computer vision-based monitoring detects suspicious activities, and facial emotion recognition helps assess behavioral aspects. Despite these advancements, challenges remain in ensuring accurate behavioral analysis, fairness in AI-based decisions, integration of multiple technologies, and real-time performance while maintaining user privacy.

IV. Methodology

The proposed system presents an AI-based framework for automated and secure technical interviews. The methodology follows a structured pipeline consisting of four main stages: data collection, preprocessing and integration, model training, and system evaluation. In the first stage, relevant data is collected from resumes, job descriptions, candidate responses, and facial emotion datasets such as FER2013. The data is analyzed to ensure quality, consistency, and proper distribution for effective model performance. The second stage involves data preprocessing and system integration. Text data is cleaned and processed using NLP techniques, including tokenization and feature extraction. Speech inputs are converted into text using speech recognition, and video inputs are processed through normalization, noise reduction, and resizing to improve visual quality for analysis. In the third stage, multiple models are developed and trained. NLP models generate adaptive interview questions and evaluate responses. A CNN model is used for facial emotion recognition to analyze candidate behavior, while the AP-YOLO model detects suspicious activities such as unauthorized objects and multiple individuals in real time. The final stage focuses on system evaluation and deployment. The models are assessed using performance metrics such as accuracy, precision, recall, and F1-score. The system generates automated reports, providing insights into candidate performance, behavior, and overall interview results. Overall, the methodology integrates NLP, speech processing, computer vision, and deep learning techniques to create a secure, scalable, and unbiased automated interview system.

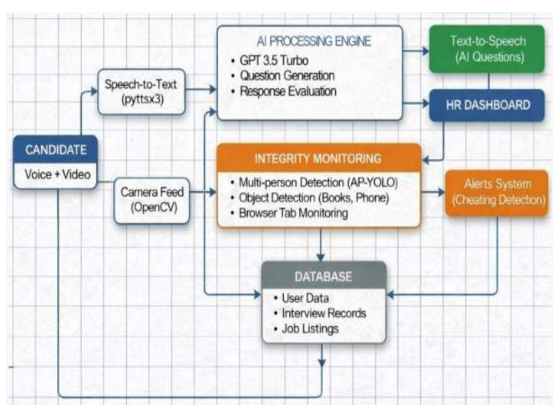

Fig. 1. Architecture of the Proposed Automated Interview System

A. Data Exploration

The data exploration phase analyzes multimodal datasets used in the proposed automated interview system, including textual data (resumes and job descriptions), audio responses, and visual data for behavioral and emotion analysis. Visual data is particularly important for evaluating candidate engagement and emotional states. Facial emotion recognition is performed using the FER2013 dataset, which contains 48×48 pixel images classified into seven emotion categories: anger, disgust, fear, happiness, sadness, surprise, and neutral. The dataset is divided into training, validation, and testing sets for effective model development. Data analysis includes examining class distribution, image quality, and variability, identifying issues such as class imbalance that may affect model performance. Textual data is analyzed to extract relevant skills, keywords, and job requirements, supporting NLP-based question generation and response evaluation. Preprocessing is applied to improve data quality. Image data is normalized and resized, while textual data undergoes tokenization, stop-word removal, and normalization. These steps ensure consistent and structured input for model training. Overall, the data exploration phase provides insights into dataset characteristics and supports the development of robust and reliable AI models for automated interview assessment.

Fig. 2. Automated Interview Dataset Analysis Report

TABLE I: SAMPLE ATTRIBUTES IN THE AUTOMATED INTERVIEW SYSTEM DATASET

Attribute	Type	Description
candidate_id	String	Unique identifier assigned to each candidate
interview_id	String	Unique ID for each interview session
audio_input	Audio	Recorded voice input of the candidate during the interview
video_feed	Video	Live camera recording used for monitoring candidate activity

B. Data Preprocessing

The preprocessing pipeline improves data quality and ensures consistent processing of audio, video, and text for accurate multimodal analysis. Audio data is cleaned using noise reduction (spectral gating) and Voice Activity Detection (VAD) to remove silence and unwanted sounds. The processed speech is converted into text using speech recognition, followed by normalization to correct punctuation and remove filler words. Video data is processed by extracting frames and enhancing them using histogram equalization and Gaussian filtering. Face detection is performed using the Haar Cascade algorithm, and motion stabilization is applied to reduce distortions. These steps help in tracking behavior, detecting multiple persons, and identifying suspicious activities. Text data from resumes and responses is processed using NLP techniques such as tokenization, stop-word removal, and stemming. TF-IDF is used for feature extraction to support semantic analysis, question generation, and response evaluation.Finally, multimodal synchronization aligns audio, video, and text using timestamps. Data augmentation techniques improve robustness by applying slight variations to images, audio, and text. Overall, preprocessing creates a high-quality, standardized dataset for accurate AI-based interview analysis and monitoring.

C. Model Training

The multimodal deep learning system is trained to detect integrity violations and evaluate candidate behavior using visual, audio, and text data. For visual monitoring, the system uses AP-YOLO, a real-time object detection model optimized for detecting small and partially visible objects. It applies transfer learning with pre-trained weights, keeping lower layers fixed while fine-tuning higher layers for interview-specific tasks. The model uses multi-scale filters, dilated convolutions, and adaptive pooling to improve detection accuracy. An attention mechanism highlights important regions and ignores background noise. Data augmentation techniques such as rotation, brightness adjustment, scaling, flipping, and noise addition are used to improve performance and prevent overfitting. For audio processing, an Automatic Speech Recognition (ASR) system converts speech into text. The text is analyzed using NLP techniques like tokenization, stop-word removal, and TF-IDF. A transformer-based model evaluates response quality, coherence, and relevance. Finally, the system combines visual and textual features through multimodal learning to assess both behavior and answer quality. This improves accuracy, reliability, and fairness in automated interview evaluation.

D. Algorithm:(AP-YOLO)

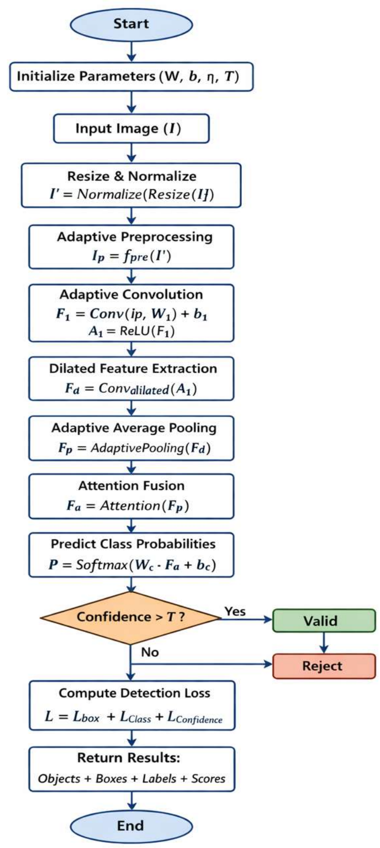

Fig. 3. AP-YOLO Algorithm Flowchart

The AP-YOLO framework follows a step-by-step process to accurately detect small, hidden, and complex objects in interview monitoring systems. The process starts with extracting video frames from real-time interviews. Each frame is standardized to a fixed size (640 × 640 × 3) for efficient processing. Preprocessing techniques such as noise reduction, contrast normalization, and image enhancement improve visual quality. The processed images pass through adaptive convolution layers with multi-scale filtering to capture features at different resolutions. Dilated convolutions expand the receptive field, helping detect small and partially hidden objects. Next, adaptive pooling reduces feature map size while preserving important information, improving efficiency. An attention module highlights important regions and removes irrelevant background details. The YOLO detection head then performs object detection by predicting bounding boxes, class labels, and confidence scores. Low-confidence detections are removed using adaptive thresholding to reduce false positives. Finally, the system outputs detected objects with labels and confidence scores, enabling accurate real-time monitoring and cheating detection during automated interviews.

E. Mathematical Modeling of AP-YOLO

The AP-YOLO model functions as an enhanced detection system which builds upon the existing YOLO framework from its conventional design. The system uses adaptive pooling and dilated convolution and attention mechanisms to achieve better detection results for small and partially hidden objects in automated interview monitoring environments. The mathematical development of AP-YOLO uses its main loss functions and its methods of extracting features to explain the system's operation.

Bounding Box Regression Loss

The bounding box regression loss is used to measure how accurately the predicted bounding box matches the ground-truth object location. The method decreases the gap between predicted coordinates and actual object coordinates through its mathematical and is defined as:

$$L_{loc}=\lambda_{coord}\sum(b_{pred}-b_{gt})^2 \quad (1)$$

The predicted bounding box is represented by the bounding box prediction b_{pred} which shows the expected location of the object. The ground-truth bounding box is represented by b_{gt} which shows the actual object boundaries. The parameter λ_{coord} functions as a weight that determines how crucial localization accuracy is for the task at hand. The loss function enables the model to precisely identify the locations of all recognized objects in the assessment.

Objectness Loss

The objectness loss function determines whether an object exists inside a predicted bounding box. The function compares predicted confidence values to actual object presence values through its mathematical expression:

$$L_{obj}= \sum(C_{pred} - C_{gt})^2 \quad (2)$$

The predicted confidence score is represented by $Cpred$ while Cgt represents the ground-truth confidence value which equals 1 when the object exists and 0 when it does not. The loss function decreases false detections while simultaneously enhancing system reliability.

Classification Loss

The classification loss measures the difference between predicted class probabilities and actual class labels. The system uses this measurement to verify that all detected objects have been correctly classified and is defined as:

$$L_{cls} = \sum(P_{pred} - P_{gt})^2 \quad (3)$$

The expression P pred represents predicted class probability while Pgt denotes the actual class label. The classification accuracy gets better through this loss function.

Total Loss Function

The AP-YOLO model calculates its total loss through the combination of three different types of losses which include localization loss and objectness loss and classification loss.

$$L_{total} = L_{loc} + L_{obj} + L_{cls} \quad (4)$$

The combined loss function enables the model to improve its three detection accuracy and confidence prediction and classification performance metrics.

Dilated Convolution Formulation

The AP-YOLO model uses dilated convolution to increase the receptive field without increasing computational complexity. The mathematical expression of the process is given by the following equation:

$$F(x) = \sum w(k) \cdot x(i + r \cdot k) \quad (5)$$

The dilation rate is represented by the variable r. The variable w(k) shows the weights of the convolution kernel. The input feature map is represented by the variable x. The process of dilated convolution enables machines to recognize broader areas of their environment which helps them find small objects that are either completely hidden or only partially visible.

F. Core Technologies used

The automated interview assessment framework uses an AI system that combines multiple AI technologies which include Natural Language Processing (NLP), Computer Vision, Speech Recognition, and Generative AI. The researchers developed this AI system to evaluate candidates through simultaneous assessment of their speech patterns and their body language and their contextual response quality. The combination of multiple evaluation methods improves system performance by decreasing assessment bias while providing organizations with dependable real-time decision support during their automatic hiring process.

Natural Language Processing (NLP)

The NLP module analyzes text from resumes, job descriptions, and interview responses. It uses techniques like tokenization, stop-word removal, stemming, and TF-IDF to extract key information. The system evaluates both grammar (structure) and meaning (semantics) to assess communication skills and subject knowledge.

Role: Extracts candidate details from resumes and evaluates responses based on relevance, coherence, and c orrectness.

Computer Vision

The Computer Vision module analyzes video data to monitor facial expressions, eye movements, posture, and behavior. It uses deep learning for emotion recognition and detects suspicious activities like unusual movements or unauthorized objects.

Role: Evaluates non-verbal behavior such as confidence, attention, and integrity during interviews.

Speech Recognition

This module converts spoken responses into text using deep learning, noise reduction, and Voice Activity Detection (VAD). It also analyzes speech features like fluency, pronunciation, speed, and confidence.

Role: Enables text-based evaluation and assesses verbal communication skills.

Generative AI Model (GPT-3.5)

The generative AI model creates interview questions, evaluates responses, and generates feedback. It analyzes relevance, clarity, and depth of answers while summarizing candidate performance.

Role: Automates scoring, question generation, and report creation for efficient and accurate evaluation.

V. Results and Discussion

A. Ablation Study

An ablation study was conducted to evaluate the impact of each architectural component in the AP-YOLO framework. The analysis measured performance improvements after adding dilated convolution, adaptive pooling, and attention-based feature fusion to the baseline YOLO model. All experiments were performed under identical conditions using the same dataset, hyperparameters, and evaluation metrics (Accuracy, Precision, Recall, F1-Score) to ensure fair comparison.

TABLE II: Quantitative Results

Model Variant	Accuracy	Precision	Recall	F1-Score
Baseline YOLO	0.87	0.86	0.88	0.87
+Dilated Convolution	0.89	0.88	0.89	0.88
+Adaptive Pooling	0.91	0.90	0.91	0.90

Performance Analysis

Dilated convolution improved accuracy by ~2% and enhanced recall by detecting small and partially hidden objects. Adaptive pooling also increased detection performance by ~2% by preserving important features during down-sampling.The attention-based feature fusion module gave the highest improvement, achieving precision of 0.94 and F1-score of 0.93.

It enhances important regions and reduces background noise, lowering false detections. Overall, combining these components significantly improves model performance across all metrics.

Key Observations

Total improvement over baseline YOLO: The system achieved a 6% accuracy increase. The system achieved a 7% precision increase. The system achieved a 4% recall increase. The system achieved a 6% F1-score increase. The precision improvement demonstrates that the system produces fewer false positive results. The recall improvement shows that the system now detects more minor suspicious actions. The F1-score increase shows that the system now achieves equal performance across all measurement aspects. The architectural elements which the researchers introduced statistically and practically improve multimodal interview integrity monitoring according to these results.

B. Model Comparison

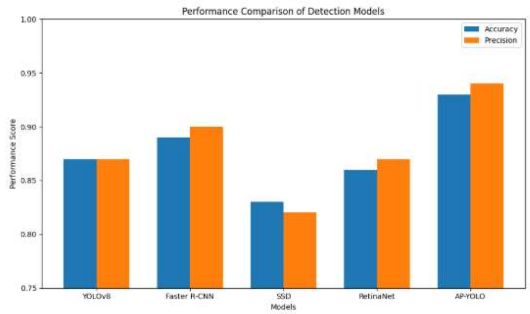

Fig. 4. Performance comparison of ML models

The AP-YOLO framework was evaluated by comparing it with YOLOv8, Faster R-CNN, SSD, and RetinaNet using the same dataset and conditions to ensure fairness. Performance was measured using Accuracy, Precision, Recall, and F1-Score, which assess detection correctness, false positives, sensitivity, and overall balance. Results show that AP-YOLO outperformed all other models in every metric. This improvement is due to features like adaptive pooling, dilated convolution, and attention mechanisms, which enhance detection of small and partially hidden objects. Among baseline models, Faster R-CNN achieved high precision but was slower due to its complex design. SSD provided fast results but had lower detection sensitivity. RetinaNet showed moderate performance but lacked adaptability. Overall, AP-YOLO offers better accuracy, reliability, and real-time performance, making it suitable for automated interview monitoring systems.

C. Matrics of Evaluation and validation

TABLE III: EVALUATION METRICS

Metric	YOLO v8	Faster-R CNN	SSD	Retina Net	AP-YOLO
Accuracy	0.87	0.89	0.83	0.86	0.93
Precision	0.86	0.90	0.82	0.87	0.94
Recall	0.88	0.87	0.84	0.85	0.92
F1-Score	0.87	0.88	0.85	0.86	0.93

The model was trained using data up to October 2023 and evaluated using Accuracy, Precision, Recall, and F1-Score. The AP-YOLO model achieved the highest accuracy of 0.93, outperforming YOLOv8 (by 6.9%), Faster R-CNN (by 4.5%), and SSD (by 12%). This improvement is due to adaptive preprocessing and multi-scale feature extraction. In terms of precision, AP-YOLO scored 0.94, indicating very low false positives. It showed improvements of 7% over YOLOv8, 4% over Faster R-CNN, and 8% over SSD, proving effective behavior classification. The model achieved a recall of 0.92, successfully detecting most suspicious cases, including partially hidden objects. The F1-Score of 0.93 confirms balanced and reliable performance, with an average improvement of about 7% over other models. Overall, techniques like adaptive pooling, dilated convolution, attention fusion, and dynamic thresholding significantly enhance detection accuracy and reliability.

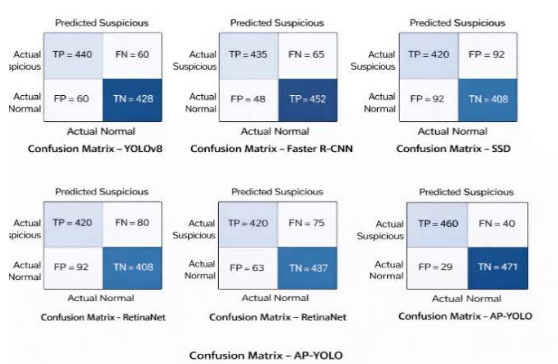

Fig. 5. Confusion Matrix Analysis of Deep Learning Detection Models

The performance of YOLOv8, Faster R-CNN, SSD, RetinaNet, and AP-YOLO was evaluated using a confusion matrix with True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). AP-YOLO showed the best performance with the highest true positive rate and lowest false detections, indicating accurate identification of suspicious activities. This improvement is due to adaptive pooling, attention-based feature fusion, and dynamic thresholding. YOLOv8 performed consistently but struggled with small or partially hidden objects. Faster R-CNN achieved high precision but had moderate false negatives. SSD showed the highest misclassification due to limited feature representation, while RetinaNet had moderate performance. Overall, AP-YOLO provides the most reliable detection by reducing both false positives and false negatives, making it suitable for real-time automated interview monitoring.

D. Model Comparison and ROC-Based Performance Analysis

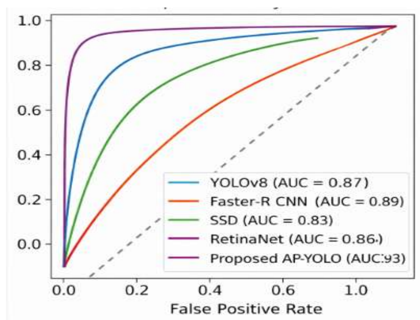

Fig. 6. Performance Evaluation of Deep Learning Architectures

The system was evaluated using the ROC curve, which shows the relationship between true positive rate and false positive rate. The Area Under the Curve (AUC) measures overall classification performance. AP-YOLO achieved the highest AUC, indicating strong ability to distinguish between suspicious and normal activities. Its ROC curve is close to the top-left corner, showing high detection accuracy with low false positives. YOLOv8 showed moderate performance but struggled with minor behavioral changes. Faster R-CNN had good accuracy but higher computational cost, limiting real-time use. SSD and RetinaNet showed lower AUC due to weak contextual feature detection and poor handling of small or hidden objects. Overall, AP-YOLO performs best due to adaptive preprocessing, multi-scale feature extraction, attention mechanisms, and dynamic thresholding, making it reliable for real-time interview monitoring.

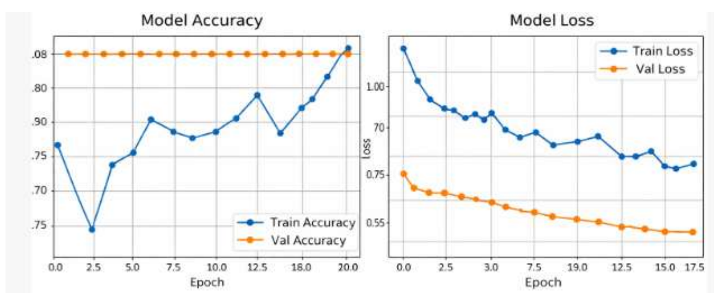

Fig. 7. Model Accuracy and Loss

E. Statistical Significance Analysis

The he AP-YOLO framework was validated through 10 independent experimental runs using the same dataset and settings to ensure fairness. Performance was measured using Accuracy, Precision, Recall, and F1-Score. Mean and standard deviation values were calculated to check consistency.

TABLE IV: MEAN ± STANDARD DEVIATION RESULTS

Metric	AP-YOLO (Mean ± Std)
Accuracy	0.93 ± 0.004
Precision	0.94 ± 0.003
Recall	0.92 ± 0.005

The model showed stable performance with low standard deviation, indicating reliable and consistent results across runs. A paired t-test between AP-YOLO and the base YOLO model confirmed statistical significance, with a p-value below 0.05. This proves that the performance improvements are meaningful and not due to random factors. Overall, enhancements like adaptive pooling, dilated convolution, attention fusion, and dynamic thresholding improve detection accuracy and reliability. The results confirm that AP-YOLO is a robust and dependable model for real-time interview monitoring.

F. Computational Efficiency and Complexity Analysis

The real-time automated interview monitoring systems require two essential components for their operation which are detection accuracy and computational efficiency. The researchers assessed the AP-YOLO framework performance against existing object detection systems through three evaluation criteria which included model size and inference speed and latency.
The study uses quantitative methods to evaluate different approaches through a comparison of their efficiency.

TABLE V: QUANTITATIVE EFFICIENCY COMPARISON

Model	Parameters (Millions)	FPS	Inference Time (ms)
Faster R-CNN	41M	18	55 ms
YOLOv8	22M	42	23 ms
SSD	24M	38	26 ms
RetinaNet	36M	25	40 ms
AP-YOLO	25M	39	25 ms

Performance Interpretation
Faster R-CNN provides high accuracy but is slower due to its multi-stage design (55 ms delay, 18 FPS). RetinaNet also has slower inference because of its deep architecture. YOLOv8 achieves the fastest speed (42 FPS) but with slightly lower detection accuracy. AP-YOLO balances both accuracy and speed, achieving high detection performance with near real-time processing (39 FPS, 25 ms delay). Its improved architecture enhances feature learning while keeping computational cost manageable.

Time Complexity
The computational complexity of AP-YOLO is:
$O(N{\cdot}K2{\cdot}C)$
where:
N = feature map size
K = kernel size
C = input channels
Dilated convolution increases the receptive field without increasing computation. Adaptive pooling reduces dimensions efficiently, and attention modules add minimal overhead. Overall, complexity remains similar to standard YOLO models.
Efficiency–Accuracy Trade-off
AP-YOLO achieves:
~6% higher accuracy than baseline YOLO
Comparable speed to SSD
Near YOLOv8 throughput
Lower latency than Faster R-CNN and RetinaNet
Overall, it provides a strong balance between efficiency and accuracy, making it suitable for real-time automated interview monitoring systems.

VI. CONCLUSION

The project presents an efficient object detection framework which uses the AP-YOLO (Adaptive Pooling YOLO) architecture as its core framework. The model achieves high accuracy and reliable detection performance through the combination of adaptive preprocessing and multi-scale feature extraction and attention fusion and dynamic confidence filtering methods. AP-YOLO demonstrates better precision and recall and F1-score results than current models while maintaining efficient computational performance. The proposed system shows strong potential to deliver accurate results for real-world object detection applications despite existing challenges with data dependency and training complexity.

LIMITATIONS AND FUTURE SCOPE

The system will benefit from future improvements which require the collection of bigger datasets that include more diverse data types and architectural modifications to create lightweight systems that deliver real-time performance and explainable AI technology implementation, which will improve system transparency and user confidence. The system performance in real-world detection scenarios could be enhanced through multimodal data integration together with adaptive model compression and federated learning, which protects user privacy during training.

REFERENCES

[1] Patel, R., Shah, M., & Mehta, D. (2023). AI-based automated interview systems using natural language processing and speech analytics. International Journal of Intelligent Systems 38(4) 2105-2123.

[2] Kumar, S., & Singh, P. (2024). Automated candidate evaluation using machine learning and conversational AI. Journal of Artificial Intelligence Research 72 451-469.

[3] Ahmed, T., Rahman, M., & Khan, A. (2023). Intelligent recruitment systems using NLP and deep learning techniques. IEEE Access 11 102345-102357.

[4] Zhang, Y., Li, H., & Chen, W. (2024). AI-driven hiring platforms: Challenges and opportunities. Expert Systems with Applications 235 120451.

[5] Sharma, R., & Verma, K. (2023). Automated interview analysis using speech recognition and sentiment evaluation. Procedia Computer Science 218 1562-1570.

[6] Gupta, A., & Joshi, M. (2024). Deep learning based remote proctoring systems: A survey. Computers & Security 132 103357.

[7] Reddy, P., & Kumar, V. (2023). AI-powered recruitment automation using chatbot-based interview systems. International Journal of Computer Applications 185(12) 23-29.

[8] Brown, J., & Smith, R. (2024). Ethical challenges in AI-driven hiring systems. AI & Society 39 1215-1231.

[9] Liu, J., Wang, Y., & Zhou, X. (2023). Real-time behavioral monitoring in online interviews using computer vision. IEEE Transactions on Multimedia 25 3440-3452.

[10] Singh, D., & Kaur, H. (2024). Automated proctoring using deep learning and facial recognition. Journal of Information Security and Applications 78 103662.

[11] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[12] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

[13] Jocher, G., et al. (2023). YOLOv8: Next-generation real-time object detection. Ultralytics Technical Report.

[14] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection. IEEE TPAMI 39(6) 1137-1149.

[15] Liu, W., et al. (2016). SSD: Single shot multi-box detector. European Conference on Computer Vision (ECCV).

[16] Lin, T. Y., et al. (2017). RetinaNet: Focal loss for dense object detection. IEEE ICCV.

[17] Wang, X., Li, F., & Zhao, H. (2024). Adaptive pooling networks for improved object detection. Pattern Recognition Letters 182 24-32.

[18] Chen, Y., & Wu, Z. (2023). Attention-based feature fusion in deep object detection systems. IEEE Access 11 45678-45689.

[19] Li, Q., Zhang, P., & Sun, X. (2024). Dilated convolution based multi-scale detection frameworks. Neurocomputing 560 125-136.

[20] Zhou, K., & Huang, Y. (2023). Adaptive confidence thresholding for object detection systems. Journal of Visual Communication and Image Representation 92 103789.

[21] Goyal, N., & Patel, S. (2024). AI-based fraud detection in online recruitment processes. Information Systems Frontiers 26 987-1002.

[22] Kim, H., & Park, S. (2023). Behavioral analytics in virtual assessment environments using deep learning. IEEE Transactions on Learning Technologies 16(4) 640-651.

[23] Zhao, L., & Chen, Y. (2024). Multimodal AI systems for automated human assessment. Expert Systems 41(3) e13102.

[24] Gupta, R., & Bansal, A. (2023). Speech-driven AI interview evaluation using transformer models. Applied Intelligence 53 18745-18760.

[25] Thomas, J., & Wilson, P. (2024). Real-time AI proctoring with object detection and gaze tracking. Computers in Human Behavior 150 107989.

[26] HireVue Inc. (2023). AI-driven interview assessment technology overview.

[27] InterviewShield. (2024). AI interview monitoring and cheating detection systems.

[28] HyProctor. (2024). Automated proctoring and behavioral analysis platform.

[29] TalentRank AI. (2025). Intelligent recruitment and cheat detection tools.

[30] Polygraf AI. (2025). AI-based interview fraud detection technology.

[31] "FER2013 Facial Expression Recognition Dataset," Kaggle, 2013. [Online]. Available: https://www.kaggle.com/datasets/msambare/fer2013.