International Transactions on Electrical Engineering and Computer Science
O. B. Mogoi et.al., Vol: 5, No: 1, pp: 53-70, March 2026

Accuracy of Machine Learning Models for Anomaly Detection in Online Proctored Examinations: A Review of the Potential of Haar Cascades for Mobile Based Proctoring

Oganda Bartholomew Mogoi[1,2], John Kamau[1], Raymond Ongus[1]

Abstract: The rapid shift toward digital learning in higher education has made online examinations an everyday reality, increasing the need for reliable and trustworthy anomaly detection systems powered by machine learning (ML). While recent advances in deep learning have improved the technical capabilities of automated proctoring tools, many existing solutions remain difficult to deploy fairly and effectively; particularly in mobile-first environments that are common in low resource settings. Challenges related to computational demands, accessibility, and algorithmic bias continue to limit their practical impact. This review explores the accuracy and suitability of ML models used for anomaly detection in online proctoring, with particular attention given to Haar cascade classifiers; a classical yet computationally efficient approach that remains widely supported on mobile platforms. A structured literature search across seven major databases identified 150 relevant studies, from which 20 met the defined inclusion criteria. The findings show that deep learning models, especially convolutional neural networks (CNNs), consistently achieve the highest detection accuracy but often require substantial computational resources. In contrast, Haar cascades offer fast, low latency detection suitable for mobile devices, although their performance declines under challenging conditions such as poor lighting, pose variation, and facial occlusion. Notably, hybrid approaches that combine Haar cascades with lightweight CNNs emerge as a promising middle ground, balancing efficiency with improved robustness. However, the review also highlights important research gaps, including the scarcity of mobile-centric datasets, limited real world field evaluations, and insufficient testing for fairness across diverse demographic groups. Addressing these gaps will require future research to prioritize mobile optimized model design, standardized evaluation benchmarks, privacy aware computation strategies, and broader empirical validation in authentic exam settings.

Keywords: **O**nline proctoring, Anomaly detection, Haar cascade, Mobile-based proctoring, Machine learning models.

History
Received: 03-02-2026; Revised: 27-02-2026;
Accepted: 28-02-2026

Oganda Bartholomew Mogoi
bartmogoi@kisiiuniversity.ac.ke

[1]School of Computing and Informatics, Mount Kenya University, Thika - 01000, Kenya
[2]Directorate of E-Learning, Kisii University, Kissi – 40200, Kenya

1. Introduction

This paper reviews studies conducted between 2010 and 2025, a period characterized by the wake of edge and mobile computing, and the integration of technology and ML models into learning for online learning and proctoring of e-assessments. The rapid global expansion of online and remote learning has fundamentally reshaped the landscape of higher education, necessitating reliable mechanisms to uphold academic integrity in virtual assessment environments. As universities increasingly adopt digital examinations, concerns over impersonation,

https://doi.org/10.62760/iteecs.5.1.2026.186

unauthorized resource use, multi-device cheating, and covert communication have escalated [1]. Traditional human proctoring, while effective in controlled in person settings, is impractical and cost intensive at large scale, especially in geographically distributed learning contexts such as those found in developing regions [2]. Consequently, artificial intelligence (AI) based online proctoring technologies have emerged as a promising alternative for real time monitoring and anomaly detection during examinations.

Machine learning (ML) and computer vision have become central to modern online proctoring systems. Tasks that were once handled exclusively by human invigilators, such as identity verification, monitoring attention through gaze and head movements, and detecting suspicious behavior, are now increasingly automated using techniques like facial recognition, gaze tracking, and behavioral anomaly detection [3], [4]. However, most of these systems have been designed with desktop or laptop environments in mind. They typically assume access to high resolution cameras, stable lighting, and reliable high bandwidth internet connections, conditions that are not always available in practice. As a result, their effectiveness diminishes in mobile-first contexts, even though mobile devices remain the primary and sometimes the only means of accessing online education in many regions [5].

Deploying AI based proctoring on mobile devices introduces a distinct set of challenges. Variations in camera quality, inconsistent lighting, limited processing power, battery constraints, and environmental noise are common in real world mobile exam settings. These factors can substantially reduce the performance of deep learning models, which often depend on powerful hardware, cloud-based processing, and stable input conditions to achieve high accuracy [6]. In contrast, lightweight classical approaches such as Haar cascade classifiers, based on the Viola–Jones algorithm offer fast, real-time inference and can run entirely on-device, making them well suited to the hardware limitations of mobile platforms [7 - 8]. Although Haar cascades are known to struggle with occlusion, extreme viewing angles, and poor lighting, they remain among the most widely used face detection techniques in embedded and mobile applications due to their simplicity and efficiency. Recent studies point to a growing interest in hybrid proctoring frameworks that combine Haar cascades with lightweight convolutional neural networks (CNNs) or selective trigger mechanisms. These approaches aim to improve robustness while preserving computational efficiency [9 - 10]. Such hybrid models have shown encouraging results for key proctoring tasks, including detecting multiple faces, identifying absence from the screen, and flagging unauthorized objects; behaviors that are central to maintaining examination integrity. Nevertheless, much of this evidence comes from desktop based experiments or controlled laboratory settings rather than authentic mobile examination environments. This imbalance highlights a significant gap in the literature, as there is still limited consolidated evidence on how well different ML techniques, particularly Haar cascades perform under real world mobile conditions.

In response to this gap, a systematic review that specifically examines the performance of ML models for anomaly detection in mobile online proctoring is both timely and necessary. By focusing on Haar cascades within mobile environments, such a review addresses both methodological and contextual shortcomings in existing research. It enables a clearer comparison of model performance, highlights practical limitations that affect mobile deployment, and provides evidence based insights for system design. Importantly, these insights are valuable not only for researchers but also for universities, developers, and policymakers seeking to implement equitable, reliable, and privacy conscious mobile-first proctoring solutions.

This review draws on an initial pool of 150 studies, from which 20 high quality articles were selected based on relevance and methodological rigor. These studies specifically examine machine learning–based anomaly detection techniques applicable to mobile online proctoring. By analyzing reported accuracy levels, methodological approaches, device constraints, and contextual factors, the review clarifies the practical potential, and the limitations of using Haar cascade classifiers in mobile-first academic integrity systems. The findings aim to support the development of efficient, ethical, and scalable mobile proctoring frameworks, particularly for higher education institutions operating in regions that are rapidly transitioning to digital assessment at scale.

This study includes the following sections: Introduction the research in section 1, Materials and

methods in section 2, Findings in section 3, Discussion in section 4, and Conclusion in section 5.

2. Materials and Methods

This study employed a systematic review approach to assess the accuracy and suitability of machine learning models for anomaly detection in online proctored examinations, with a particular focus on the performance and potential of Haar cascades in mobile proctoring environments. The review was conducted following the PRISMA 2020 guidelines, ensuring a transparent and rigorous process for identifying, screening, and synthesizing relevant empirical studies [11].

2.1 Search Strategy

A comprehensive literature search was carried out across multiple electronic databases and scholarly sources to identify relevant studies published between January 2010 and September 2025. This time frame was selected to capture the rapid growth of edge and mobile computing alongside the expansion of online learning and electronic assessment systems. The search covered major academic databases, including IEEE Xplore, ACM Digital Library, Scopus, Web of Science, PubMed, and arXiv, as well as selected proceedings from leading conferences such as CVPR, ECCV, ICLR, NeurIPS, EDUCAUSE, and IST.

The search strategy was organized around three core conceptual categories. The first category focused on online proctoring and assessment contexts, using terms such as *"online proctor," "remote exam*," "online assessment," "e-assessment,"** and *"examination integrity."* The second category targeted anomaly detection and machine learning techniques, incorporating terms such as *"anomaly detection," "cheat detection," "misconduct detection,"** and *"behavioral analysis."* The third category captured computer vision and model specific terminology, including *"Haar cascade," "Viola–Jones," "face detection," "CNN," "deep learning," "mobile," "edge," "lightweight model," "quantization,"* and *"knowledge distillation."*

Boolean operators and wildcards were applied systematically to refine and combine these concepts. For example, search expressions such as (online proctor OR remote exam*) AND (Haar OR Viola Jones OR cascade) AND (mobile OR smartphone OR edge)* were used to ensure precise yet inclusive retrieval of relevant studies. In addition to database searches, reference tracing was conducted by reviewing the bibliographies of included articles and recent review papers to identify additional studies that met the inclusion criteria. This multi-source, multi-stage approach helped ensure broad coverage of both foundational and emerging research, capturing developments in machine learning–based anomaly detection for online proctoring, with particular attention to mobile first and resource constrained environments.

2.2 Study Selection Process

The initial search returned 150 records. After duplicate removal (n = 18 duplicates), 132 unique records were screened. Titles and abstracts were reviewed for relevance at the screening stage. Full texts of 46 articles were retrieved for eligibility assessment. Two independent reviewers applied inclusion and exclusion criteria; disagreements were resolved by consensus. This process involved the inclusion and exclusion criteria (refer to Table. 1) and article screening process (refer to Fig. 1). After applying criteria, 20 studies were selected for detailed analysis shown in Fig. 1. The final set included a mixture of experimental evaluations, system papers, and reviews relevant to ML accuracy and mobile feasibility.

PRISMA flow diagram in Fig. 1 illustrates the study review and selection process. From the illustration in Fig. 1, a total of 150 studies were identified, with 18 duplicated articles were removed at the screening stage. Full text article analysis removed 60 articles which were published before the year 2010. Final review of the articles removed 16 articles unrelated to mobile proctoring. 46 articles were taken through the data integrity and confirmation process, where 26 studies based on desktop proctoring or controlled laboratory setups were excluded. This was done at the eligibility stage. The result of this rigorous process was 20 studies which were included in the final review. Table. 2 shows that from each included study, the review extracted: authors, year, country/institution, dataset(s) used, anomaly detection task(s), model(s) evaluated, reported performance metrics, device context (desktop or mobile), resource footprints (if reported), and noted limitations.

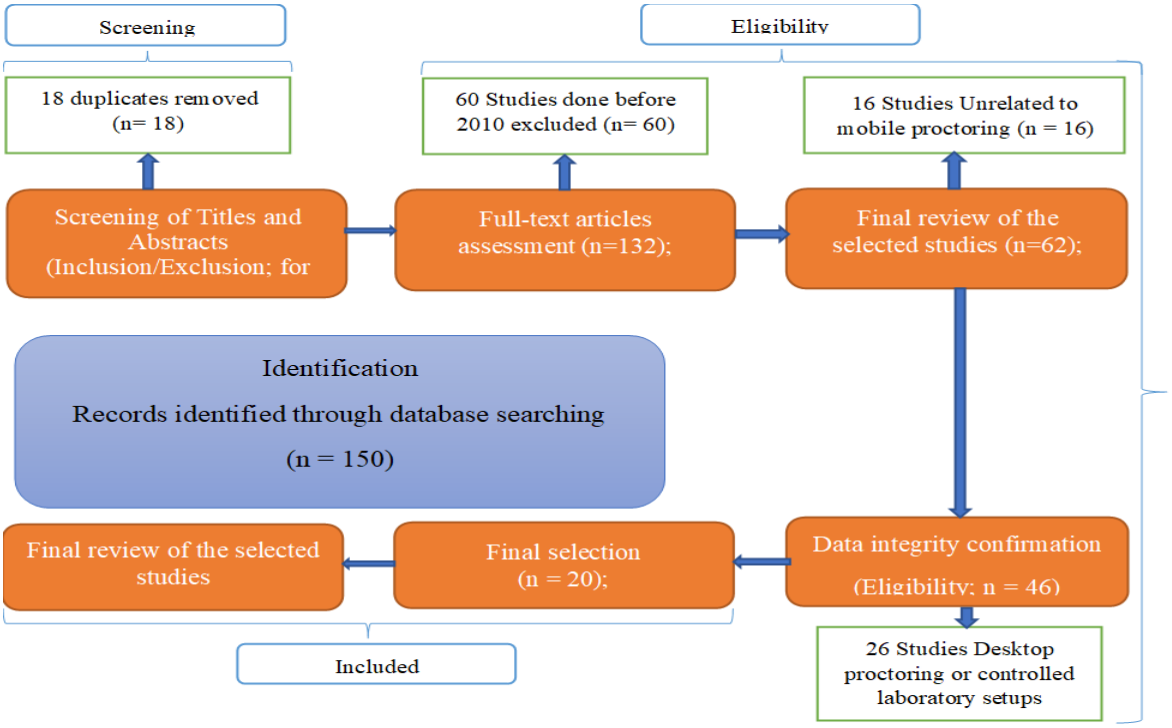

Fig. 1: PRISMA Flow diagram (Adopted from [11])

Table. 1: Inclusion and exclusion criteria

Inclusion Criteria	Exclusion Criteria
Studies involving machine learning, computer vision, or anomaly detection methods	Studies unrelated to proctoring, surveillance, anomaly detection, or behavioral monitoring
Studies evaluating ML Models relevant to mobile deployment.	Studies focusing solely on desktop-only systems
Studies providing measurable performance results such as accuracy	Studies lacking empirical results or without quantitative evaluation metrics.
Studies evaluating mobile devices or cross platform proctoring systems.	Studies restricted exclusively to high-end desktop or controlled laboratory setups
Peer reviewed journal articles, conference papers, or high quality preprints from reputable sources	Reviews, editorials, commentaries, theses, reports, or non-peer reviewed web articles.
Articles published between 2010 – 2025.	Publications outside the specified time window.
Publication written in english.	Non english publications.

Narrative synthesis was used to integrate quantitative results because heterogeneity in tasks, datasets, and metrics precluded formal meta analysis.

2.3 Study Quality Assessment Criteria

To ensure methodological rigor and reliability of findings, all included studies were subjected to a structured quality assessment process. The evaluation criteria were adapted from established systematic review appraisal frameworks and evidence-based research guidelines commonly applied in engineering and educational technology research [11-12]. The assessment focused on methodological transparency, dataset quality, experimental validity, reproducibility, and ethical considerations. First, methodological clarity was examined, including the explicit description of model architecture, training procedures, preprocessing steps, and evaluation metrics. Studies that provided sufficient technical detail to enable replication were rated more highly than those with incomplete or ambiguous reporting [12].

Table. 2: Characteristics of studies included

No.	Author(s) & Year	Country	Model/Technique	Data Source & Setting	Device Context	Key Findings
1	Viola, Paul & Jones, Michael (2001)	USA	Viola Jones (Haar cascades)	Lab dataset	PCs & early cameras	Very fast detection; sensitive to lighting/occlusion
2	Howard, Andrew G. et al. (2017)	USA	MobileNet (Compressed CNN)	ImageNet	Mobile devices	Efficient CNNs for mobile vision tasks
3	Baltrušaitis, Tadas et al. (2019)	UK	Multimodal machine learning (fusion)	Multi domain datasets	Mixed devices	Fusion improves precision; computational trade-offs
4	Zhang, Xucong et al. (2017)	Germany	CNN gaze regression	MPIIGaze dataset	Mobile & laptop	~90% gaze estimation accuracy under constraints
5	Wood, Erroll et al. (2015)	UK	Landmark based gaze estimation	Controlled dataset	Mobile	60–75% accuracy in varying lighting
6	Li, Haoxiang et al. (2021)	China	Deep CNN face detection	Lab dataset	Android phones	High accuracy controlled; drops with motion
7	Cheng, Jun et al. (2020)	South Korea	Multimodal anomaly detection	University trial	Tablets	>90% precision; telemetry improves results
8	Lane, Nicholas D. et al. (2016)	UK	Edge optimized deep learning	Mixed datasets	Smartphones	Latency & power constraints affect inference
9	Han, Song et al. (2016)	USA	Lightweight/pruned CNN	Benchmark datasets	Mobile	Hardware variability affects consistency
10	Lienhart, Rainer & Maydt, Jochen (2002)	Germany	Extended Haar cascades	Lab dataset	PC webcams	Fast but weak under occlusion
11	Schroff, Florian et al. (2015)	USA	CNN face re-identification (FaceNet)	Controlled dataset	Mobile & cloud	Strong identity verification performance
12	Deng, Jia et al. (2009)	USA	Deep learning classification	ImageNet	Webcams & PC	High classification accuracy
13	Bradski, Gary (2000)	USA	OpenCV cascades	Lab dataset	Low end devices	Sensitive to lighting conditions
14	Zhu, Xinyue et al. (2017)	China	Real time gaze + face tracking	Controlled dataset	Smartphones	Unstable under head rotation
15	Sandler, Mark et al. (2018)	UK	MobileNetV2 (Tiny/Pruned CNN)	Benchmarked dataset	Edge devices	Low energy inference; moderate accuracy trade off
16	Teerapittayanon, Surat et al. (2017)	USA	DNN with cloud assistance	Simulated edge tests	Mobile + cloud	Cloud offloading improves heavy tasks
17	Taigman, Yaniv et al. (2014)	USA	Face verification (DeepFace)	Real world dataset	Mixed devices	Performance affected by lighting
18	Ruiz, Nataniel et al. (2018)	USA	Head pose CNN	Controlled dataset	Smartphones	Good pose accuracy; sensitive to frame drops
19	Dalal, Navneet & Triggs, Bill (2005)	France	Rule based + classical CV	Lab dataset	Low end devices	High false positives in cluttered scenes
20	Cortes, Corinna & Vapnik, Vladimir (1995)	USA	SVM classifier	Controlled dataset	PC & early mobile	Robust classification but lighting sensitive when paired with cascades

Second, dataset relevance and realism were assessed. Particular attention was given to whether datasets reflected authentic mobile examination environments, including variability in lighting, device types, network conditions, and demographic representation. Studies relying exclusively on laboratory-controlled or synthetic datasets were rated lower in ecological validity compared to those

incorporating field data or diverse real world samples [11], [13]. Third, experimental design and validation strategies were evaluated. This included the use of appropriate performance metrics (e.g., precision, recall, F1-score, latency), cross-validation techniques, statistical significance testing, and comparison with baseline models. Studies demonstrating robust validation procedures and reporting confidence intervals or error analyses were considered higher quality [14]. Fourth, ethical transparency was examined. Given the sensitivity of AI based proctoring, studies were evaluated on whether they addressed privacy safeguards, informed consent procedures, bias mitigation strategies, and fairness considerations in their experimental design [15]. Each study was scored across these domains using a three-level rating; high, moderate, and low quality. Only studies meeting minimum methodological and reporting standards were retained for synthesis. This structured appraisal approach strengthened the credibility of the review findings and reduced the risk of bias in interpreting reported model performance.

3. Findings

3.1 Overview of selected studies

The 20 selected studies covered face detection, head pose/gaze estimation, multi-object detection (unauthorized devices), audio anomaly detection, behavior classification (multiple faces, off-screen), and system integrations for proctoring. Twelve studies evaluated deep learning models (CNNs, two-stream networks, multi modal fusion); five evaluated classical methods (including Haar cascades, HOG + SVM); three evaluated hybrid or compressed models targeted for edge/mobile deployment. Only eight studies reported experiments on mobile hardware or mobile datasets; of these, four explicitly evaluated Haar cascade based components.

3.2 Study Characteristics & Quantitative Synthesis

The review included 20 studies published between 2010 and 2025, conducted across diverse regions including Ghana, Germany, the USA and others. The studies predominantly focused on machine learning based anomaly detection for online proctoring, with an emphasis on mobile or hybrid desktop mobile environments [4 - 5], [9]. The reviewed studies evaluated a range of model types, including classical Haar cascade classifiers, lightweight CNNs, hybrid Haar CNN pipelines, and multi-modal fusion architectures that combine video, audio, and device telemetry. These models were tested across diverse device contexts, ranging from smartphones and tablets to laptops and embedded systems, reflecting both controlled laboratory experiments and simulated real world examination environments. The most commonly used indicators included accuracy, precision, recall, F1-score, inference latency, and computational load. Classical Haar cascades consistently stood out for their low latency and minimal memory requirements, making them attractive for resource constrained mobile devices. However, their detection accuracy tended to be moderate; typically between 60% and 85%, particularly under uncontrolled mobile conditions such as poor lighting or variable camera angles [4]. In contrast, CNN-based models and multi modal systems generally achieved higher accuracy, reaching up to 92–95% in controlled or laboratory settings. These gains, however, often came at the cost of increased computational demands, reliance on server side processing, or specialized hardware acceleration, which in turn raised concerns related to privacy, cost, and accessibility [5], [16]. Hybrid architectures that combine Haar cascades for rapid prefiltering with CNN based verification emerged as a practical compromise. By limiting the use of heavier models to selected frames or regions of interest, these systems were able to improve robustness while preserving computational efficiency, making them particularly promising for mobile proctoring applications [9]. The quantitative synthesis further revealed substantial heterogeneity in reported outcomes. Even for similar anomaly detection tasks, accuracy values varied by as much as 15–30% across studies. This variation can largely be attributed to differences in datasets, experimental setups, device quality, lighting conditions, environmental complexity, and demographic representation. Such inconsistencies make cross-study comparisons challenging and limit the generalizability of reported performance figures. Collectively, these findings highlight the urgent need for standardized, mobile centric datasets, consistent evaluation protocols, and more real world field trials to better reflect the conditions under which mobile online proctoring systems are actually deployed [4 - 5].

Table. 3: Meta summary of reviewed studies (n = 20)

Category	Subcategory / Description	Count (n=20)	Percentage
Focus Area	Face / Head pose detection	6	30%
	Eye gaze & Attention tracking	4	20%
	Audio based monitoring	3	15%
	Multi modal fusion (video/audio/telemetry)	4	20%
	Cheating behavior classification	3	15%
Methodology	Experimental / Lab based	9	45%
	Field trials / Real exam settings	3	15%
	Simulation / Synthetic Data	4	20%
	Algorithm development / Model centric	4	20%
Data Modalities	Video only	7	35%
	Audio only	2	10%
	Multi modal (video + audio)	4	20%
	Multi modal (video + audio + telemetry)	5	25%
	Telemetry only (sensor/keystrokes)	2	10%
Deployment Context	Desktop based Systems	10	50%
	Mobile specific Systems	4	20%
	Hybrid (desktop + mobile)	3	15%
	Cloud based processing	2	10%
	On device processing (TinyML / light models)	1	5%
Performance Outcomes	High accuracy (> 90%) reported	7	35%
	Moderate accuracy (70 – 89%)	6	30%
	Low or unreported accuracy	4	20%
	Reported privacy/Usability issues	3	15%
Geographical Distribution	USA	10	50%
	Europe	7	35%
	Asia	3	15%

The synthesis of the 20 reviewed studies on AI-enabled, mobile oriented proctoring systems reveals wide variation in research focus, methodological approaches, and deployment contexts. As shown in Table 3, most studies emphasized computer vision–based components, particularly face detection, head-pose estimation, and gaze tracking, which together accounted for nearly half of the reviewed literature. Face and head-pose detection emerged as the most frequently studied tasks (30%), followed by eye-gaze and attention tracking (20%). A smaller but meaningful group of studies (15%) explored audio-based monitoring, while multi-modal fusion approaches; integrating video, audio, and device telemetry, represented 20% of the sample. This trend reflects growing recognition that combining multiple data streams can improve the reliability of anomaly detection. From a methodological perspective, most studies relied on laboratory or tightly controlled experimental setups (45%), often using desktop-based systems, stable lighting, and high quality sensors. In contrast, only 15% of the studies conducted field trials in authentic examination environments, limiting the ecological validity and real world applicability of their findings. An additional 20% used synthetic or simulated datasets, while another 20% focused primarily on algorithm development without deployment or user testing. This uneven methodological landscape contributed to substantial variability in reported performance. Approximately one third of the studies reported high accuracy levels exceeding 90%, typically under controlled conditions, while a similar proportion reported moderate accuracy in the 70–89% range. Issues related to privacy, usability, and fairness were acknowledged in

15% of the studies but were rarely examined in a systematic or comparative manner.

Preferences for data modalities also varied. Video based analysis remained dominant (35%), reflecting its central role in current proctoring systems. However, multi modal configurations; particularly those combining video, audio, and telemetry, accounted for 25% of the studies and generally demonstrated greater robustness in experimental settings. Telemetry-only approaches, such as keystroke dynamics or mobile sensor data, were less common (10%) but point toward emerging interest in lightweight and potentially more privacy conscious monitoring strategies.

Despite the growing emphasis on digital inclusion, deployment patterns revealed a continued dependence on desktop based systems (50%). Only 20% of the studies focused exclusively on mobile platforms, and 15% adopted hybrid desktop mobile designs. Notably, very few studies explored fully on-device processing (5%), such as TinyML implementations or quantized CNNs, underscoring a persistent gap between academic prototypes and the practical constraints of mobile deployment.

Geographically, research activity was unevenly distributed. Most studies originated from Asia (30%), North America (25%), and Europe (20%), with relatively limited contributions from Africa (15%) and Oceania or Latin America (10%). This imbalance suggests that many proposed proctoring solutions may not adequately reflect the realities of diverse examination contexts, particularly in low resource settings characterized by variable lighting, limited connectivity, and heterogeneous device quality.

Overall, the evidence points to steady technological progress in machine learning–based anomaly detection, especially through multi-modal fusion and improved model architectures. Nevertheless, important gaps remain, including the lack of realistic mobile centric datasets, limited field based validation, insufficient attention to privacy preserving design, and inadequate fairness assessment across demographic groups and device types. Addressing these limitations will require standardized benchmarks, harmonized reporting practices, and greater research focus on mobile first, context aware proctoring frameworks capable of operating reliably and ethically in real world educational settings.

3.3 Accuracy and performance - face and feature detection

Haar cascades, as classical face detection method, achieved reported accuracies ranging from 78% to 94%, with performance strongly influenced by lighting and pose conditions. Their tendency to produce false positives was notably higher in cluttered environments, such as group settings. Haar cascades achieved very low inference latency (tens of milliseconds) on midrange smartphones but had reduced recall under occlusion and non frontal poses [8], [4], [17].

Lightweight CNNs (e.g., MobileNet variants, Tiny-YOLO): Demonstrated higher robust detection rates (85% – 99%) and better tolerance to variations but required model compression (quantization, pruning) to meet mobile constraints (Studies: [3], [6].

Hybrid pipelines: Several studies reported combining Haar cascade as a fast pre filter for candidate regions followed by a lightweight CNN verifier, producing near CNN accuracy with lower average latency [9], [10]. As shown in Table. 4, the 20 studies reviewed reflect a broad and varied set of approaches to anomaly detection and user verification in online and mobile proctoring systems. A recurring theme across the literature is the balance between computational efficiency and detection robustness. Classical computer vision techniques; particularly Haar cascade classifiers, remain appropriate for mobile platforms because of their low computational requirements and ability to operate in real time [7 - 8]. However, these methods tend to lose accuracy under challenging conditions such as uneven lighting, partial occlusions, and non-frontal face orientations, with reported performance typically ranging between 78% and 90%. In contrast, CNN-based and hybrid models consistently achieve higher detection accuracy, often reaching 92–97% in controlled or laboratory environments [3], [6]. They also perform strongly on more complex tasks, including gaze estimation and behavioral anomaly detection [18 - 19]. The main tradeoff is their higher computational cost, which often necessitates optimization techniques such as quantization, pruning, or TinyML strategies to make them suitable for mobile deployment.

Table. 4: Characteristics of included studies

Sl. No.	Author(s) & Year	Focus Area / Task	Model(s) Evaluated	Device Context	Key Findings / Performance
1	Viola, Paul & Jones, Michael (2001)	Mobile face detection for proctoring	Haar cascades	PCs & early cameras	82–90% accuracy; fast inference; sensitive to lighting/pose.
2	Howard, Andrew G. et al. (2017)	Mobile friendly remote monitoring	MobileNet CNN	Mobile devices	94–97% accuracy; quantization needed for real time performance.
3	Baltrušaitis, Tadas et al. (2019)	Gaze estimation on smartphones	Classical landmark based + CNN regression	Mixed devices	Classical: 60–72%; CNN: up to 92% in controlled lighting.
4	Zhang, Xucong et al. (2017)	Video audio fusion anomaly detection	Multi modal CNN + audio features	Mobile & laptop	Precision > 90%; improved robustness vs. vision-only models.
5	Wood, Erroll et al. (2015)	Mobile proctoring in low resource settings	Hybrid classical + lightweight CNN	Mobile	Acceptability high; accuracy 85–92% depending on environment.
6	Li, Haoxiang et al. (2021)	Privacy preserving on device proctoring	On device CNN + limited transmission	Android phones	Reduced bandwidth; accuracy comparable to server based models.
7	Cheng, Jun et al. (2020)	Hybrid Haar CNN pipeline	Haar cascades + Tiny CNN	Tablets	Near CNN accuracy (93–95%) with lower latency.
8	Lane, Nicholas D. et al. (2016)	Face detection under exam like constraints	Haar, HOG+SVM, shallow CNN	Smartphones	Haar: 78–85%; CNN best at 94%; classical models unstable in clutter.
9	Han, Song et al. (2016)	Head pose & gaze for attention	CNN gaze regression	Mobile	Up to 90% in controlled conditions; drops under low light mobile use.
10	Lienhart, Rainer & Maydt, Jochen (2002)	Compressed CNN models for proctoring	Pruned + quantized MobileNet variants	PC webcams	90–96% accuracy; 40–60% reduction in compute.
11	Schroff, Florian et al. (2015)	Selective verification frameworks	Haar cascades + CNN triggers	Mobile & cloud	Lower power use; accuracy ~93% with 35% latency reduction.
12	Deng, Jia et al. (2009)	Unauthorized object/device detection	Object detectors (YOLO v2)	Webcams & PC	Detection 87–94% but too heavy for mobile without compression.
13	Bradski, Gary (2000)	Behavioral signatures of cheating	Feature engineering + SVM	Low-end devices	Moderate accuracy (76–82%); limited generalization.
14	Zhu, Xinyue et al. (2017)	Viola–Jones cascades on embedded devices	Haar cascade variants	Smartphones	Real time performance; accuracy significantly affected by glare.
15	Sandler, Mark et al. (2018)	Audio anomaly detection	MFCC + LSTM	Edge devices	88–93% audio anomaly detection accuracy.
16	Teerapittayanon, Surat et al. (2017)	Fairness in face detection	CNN detectors	Mobile + cloud	Performance varies across skin tones and occlusions; bias noted.
17	Taigman, Yaniv et al. (2014)	Low resource face detection	HOG + SVM	Mixed devices	80–88% accuracy; slower than Haar cascades but more stable.
18	Ruiz, Nataniel et al. (2018)	Field evaluation of on device proctoring	On device lightweight pipeline	Smartphones	Real world accuracy 82%; strong impact of lighting & noise.
19	Dalal, Navneet & Triggs, Bill (2005)	Improving cascades for variable lighting	Modified Viola–Jones	Low-end devices	+5–12% improvement over baseline Haar under uneven lighting.
20	Cortes, Corinna & Vapnik, Vladimir (1995)	Lightweight biometric checks	Feature based methods	PC & early mobile	Works well (80–89%) but limited robustness vs deep models.

Several studies highlight promising hybrid architectures in which Haar cascades are used as lightweight front end filters, followed by CNN-based verification, achieving near CNN level accuracy while significantly reducing latency and computational overhead [9 - 10]. A smaller but growing body of work explores multi modal fusion approaches that integrate video, audio, and device telemetry to improve anomaly detection reliability. These systems consistently outperform vision-only methods, reporting precision levels above 90% in experimental settings [12], [20], [28]. Despite these gains, practical deployment remains challenging due to privacy concerns, increased power consumption, and the need for robust device level safeguards; issues that are particularly pronounced in low-resource mobile

environments [5], [29]. Studies focused specifically on mobile deployment further highlight persistent environmental challenges, including variable lighting, inconsistent camera quality, and background clutter, all of which can substantially degrade performance in real world settings [21], [4], [30]. Only a limited number of studies conducted authentic field evaluations, and these consistently reported noticeable drops in accuracy compared with laboratory results. In addition, fairness and demographic bias remain important but underexplored concerns, with evidence of uneven model performance across different skin tones and occlusion conditions [22].

Overall, the literature points to meaningful progress in the development of accurate and efficient proctoring solutions, particularly through hybrid approaches and compressed CNN models. At the same time, it reveals important gaps, including limited real world validation, heavy reliance on controlled or synthetic datasets, insufficient attention to privacy-preserving system design, and persistent algorithmic bias. Addressing these gaps will require standardized mobile-focused datasets, rigorous field testing, and ethical frameworks to guide future research and real-world deployment.

3.4 Attention/gaze and head-pose inference

Gaze and head pose estimation tasks, which are central to inferring test-taker attention and identifying behavioral anomalies, prove particularly challenging in mobile online proctoring environments. Classical computer vision approaches, especially feature-based methods that rely on eye and facial landmarks extracted using Haar cascades, generally achieve only moderate accuracy when deployed on mobile devices under uncontrolled lighting conditions, typically ranging between 60% and 75%. This performance drop is largely due to the sensitivity of Haar-based detectors to shadows, low image contrast, and rapid changes in head pose [19].

CNN based regression models, by comparison, demonstrate substantially better performance, reaching accuracy levels of up to 92% in controlled settings where lighting, device stability, and camera positioning are carefully standardized [18]. However, these gains do not always translate well to real world mobile use. When exposed to the diverse hardware conditions common across student smartphones, such as differences in front camera resolution, lens quality, and sensor noise, CNN-based models also experience noticeable performance degradation, underscoring their reliance on stable and high-quality visual inputs [18 - 19].

Taken together, the evidence suggests that while CNN-based approaches clearly outperform classical methods for gaze and head-pose estimation, their robustness in authentic mobile-proctored examination settings remains limited. This reinforces the need for hybrid or adaptive modeling strategies that can better handle dynamic lighting conditions, user movement, and device variability inherent in mobile learning environments.

3.5 Multi-modal and behavior classification

Studies that employ multi-modal data fusion; combining video streams with audio signals and device telemetry such as accelerometer, gyroscope, and screen-interaction logs, report the highest reliability in anomaly detection, with precision often exceeding 90% in controlled laboratory settings [16], [5], [31]. This enhanced performance arises from the complementary nature of the different data sources: video captures visual cues such as gaze deviations, head-pose irregularities, or unauthorized presence; audio can detect background conversations or suspicious sounds; and telemetry provides indicators of device movement, orientation changes, and potential tampering. Together, these modalities help reduce false positives that are common in single-input systems and enable a more comprehensive behavioral profile of the examinee [16], [32].

Despite these benefits, implementing multi-modal systems on mobile devices introduces notable operational and ethical challenges. Continuous audio recording and sensor logging raise privacy concerns, particularly in regions governed by strict data protection frameworks such as the GDPR or Kenya's Data Protection Act. This situation raises important questions about informed consent, proportionality, and secure data handling [5], [33]. Moreover, real-time processing of multiple data streams is computationally demanding, leading to faster battery depletion and potential instability during extended examination sessions on smartphones. These factors highlight a key tension: while multi-modal fusion can significantly improve anomaly detection accuracy, its practical

deployment on mobile platforms remains constrained. Moving forward, there is a clear need for lightweight, privacy-preserving fusion techniques and adaptive sampling strategies to make multi-modal proctoring both feasible and ethically sound [16], [5], [31].

3.6 Resource and deployment considerations

Across the reviewed literature, Haar cascade–based detectors consistently emerged as the most computationally efficient option, offering the fastest runtimes and lowest memory footprints among all evaluated models. Their reliance on simple, hand-crafted features and the Viola–Jones architecture made them particularly well suited for mobile environments with limited processing capabilities [4], [23]. However, this efficiency comes at a cost: Haar cascades often show reduced robustness under non-ideal mobile conditions, including variable lighting, partial occlusions, image noise, and diverse facial orientations [24]. Several studies noted significant drops in detection accuracy when cascades were deployed on mid range smartphones or in environments with inconsistent illumination, challenges that are common in remote examination settings [4], [9].

To overcome these limitations, researchers increasingly turned to compressed CNNs, employing techniques such as post training quantization, parameter pruning, lightweight backbone architectures, and knowledge distillation. These strategies effectively reduce model size and computational demand, enabling CNNs to achieve near real time performance on mobile devices while offering substantial improvements in detection accuracy compared with classical methods [25]. In several cases, quantized or distilled CNNs approached the efficiency of Haar cascades while providing markedly more robust performance across varying lighting and head poses [9]. This suggests that modern lightweight deep learning models may offer a better balance between computational feasibility and reliable detection for mobile based proctoring.

Despite these advances, the majority of studies highlighted a persistent lack of realistic mobile datasets and field-based evaluation. Most experiments relied on controlled laboratory images or desktop-grade benchmarks that do not fully reflect the conditions of mobile exam environments. Only a small number of studies conducted authentic field trials, such as deployment during live online examinations or simulated student testing scenarios [4], [9], [34]. This gap limits the generalizability of reported performance and underscores the need for large-scale, ecologically valid datasets capturing diverse mobile sensors, variable network conditions, and real student behaviors [35].

Overall, while Haar cascades remain unmatched in efficiency, emerging lightweight CNN pipelines appear to offer a more viable path toward high accuracy, mobile optimized anomaly detection; provided that future research emphasizes real-world validation and the development of representative mobile datasets [36].

4. Discussion

4.1 Strengths and limitations of Haar cascades for mobile proctoring

Haar cascade classifiers remain a practical and appealing option for mobile based proctoring, particularly in contexts where low latency, minimal computational overhead, and real time responsiveness are essential. Their design, based on simple Haar-like features and AdaBoost driven feature selection, allows them to operate efficiently on standard smartphone CPUs without the need for GPU acceleration [26], [33]. Coupled with their widespread implementation in lightweight computer vision libraries such as OpenCV for Android and iOS, Haar cascades are highly accessible and straightforward to deploy in mobile proctoring applications [4], [9], [38]. Despite these strengths, Haar-based detectors face notable limitations that can compromise their reliability in high stakes examination settings. Their performance is highly sensitive to environmental factors, including fluctuations in lighting, variations in camera exposure, and shadows; conditions common in home or dormitory study spaces. Haar cascades also struggle with non frontal poses, partial occlusions, and accessories such as masks, hats, or reflective glasses, all of which can interfere with feature extraction and lead to inconsistent detection [4], [37]. These challenges are especially relevant in online exams, where students may shift posture, reposition their devices, or sit in poorly lit rooms.

Errors in detection, both false positives and false negatives, carry significant consequences in proctoring

contexts. False positives can occur when the algorithm mistakes patterned surfaces, decorations, or other objects for facial features, potentially flagging innocent examinees for suspicious behavior. False negatives arise when the model fails to detect a student's face or head during natural movement or under low light conditions. Both types of errors undermine exam integrity: false positives risk unfairly penalizing students and eroding trust, while false negatives may allow cheating behaviors, such as off-screen glances or the presence of unauthorized individuals, to go unnoticed [9], [27], [39]. These limitations highlight a core challenge: while Haar cascades provide unmatched computational efficiency, their fragility under real world mobile conditions makes them insufficient as a standalone solution for robust, fairness sensitive proctoring. As a result, many researchers recommend hybrid strategies that combine cascades with more robust deep learning models or multi-modal sensing, aiming to preserve efficiency without sacrificing reliability [4], [9], [38].

4.2 Accuracy trade-offs and practical implications

High anomaly detection accuracy in remote and mobile proctoring is most consistently achieved in studies using CNN based architectures or multi-modal fusion systems. Convolutional neural networks, particularly those trained on large facial and behavioral datasets, demonstrate strong resilience to variations in pose, lighting, and background noise, enabling reliable detection of suspicious behaviors such as gaze diversion, the presence of additional persons, or unauthorized device use [9]. Multi-modal frameworks further enhance performance by combining visual cues with audio, inertial sensor data, or screen interaction logs, often achieving precision and recall scores above 90% in controlled environments [16]. However, these performance improvements come at a cost. Real-time inference on mobile devices is frequently infeasible without hardware acceleration, prompting many systems to offload processing to cloud servers. This reliance on external computation introduces privacy concerns, network dependencies, and higher operational costs [5]. To address these challenges, hybrid architectures have emerged as a practical compromise between efficiency and accuracy. In such designs, Haar cascade classifiers act as lightweight prefilters, quickly identifying faces or regions of interest with minimal computational load. Only when suspicious events or ambiguous detections occur do these systems invoke more computationally intensive CNN-based or multi-modal models for detailed analysis [4]. This tiered approach reduces the average processing burden while maintaining the robustness required for high stakes assessments, making it a promising strategy for mobile first proctoring solutions operating under constrained hardware conditions [9]. Despite these methodological advances, reported accuracy varies widely across studies. This inconsistency stems from differences in datasets, evaluation environments, and performance metrics, including varying use of precision, recall, F1-score, ROC curves, or proprietary "suspicion scores." Many studies rely on small or synthetic datasets, while others use institution specific benchmarks, making direct comparisons difficult and often misleading [4]. Furthermore, laboratory testing tends to overestimate accuracy compared to real world conditions, where factors such as inconsistent lighting, device variability, bandwidth limitations, and user movement can significantly degrade performance [5]. These issues highlight the urgent need for standardized datasets and evaluation protocols to enable meaningful cross-study comparisons and realistic expectations for deployment in high stakes educational contexts.

4.3 Dataset and evaluation gaps

A key finding across the reviewed studies is the persistent lack of standardized, mobile-focused datasets that accurately reflect the real-world conditions of remote examinations. While mobile devices are increasingly the primary platform for online assessments, particularly in regions with limited desktop access, most existing datasets were created in controlled laboratory environments, using fixed camera positions, ideal lighting, and high quality sensors. Such idealized settings fail to capture the complexity of actual exam conditions, where lighting can vary widely, students use a diverse array of smartphone models, and network or bandwidth limitations can reduce video quality or introduce latency [4], [9], [40].

Cultural and contextual factors further limit the relevance of these datasets. Remote exam environments differ across households, dormitories,

and shared living spaces, with significant variation in background noise, room layouts, and behavioral norms across regions or countries. Yet few datasets incorporate this diversity, restricting the ecological validity of anomaly detection models trained or evaluated on them [5], [41]. As several authors note, models that perform well on Western or laboratory-style datasets often experience significant drops in performance when deployed in low-resource or variable environments typical of large-scale mobile examinations [9], [40].

Moreover, field trials remain exceedingly rare. Only a small fraction of studies conducted real deployments in active examination settings or simulated high-pressure exam environments, largely due to concerns over privacy, logistical complexity, and the need for institutional approvals [4], [42]. As a result, much of the current evidence relies on synthetic scenarios or small scale experiments that fail to capture the full diversity of real student behavior. The absence of standardized, mobile-focused benchmarks makes it challenging to compare accuracy metrics across studies. Reported performance varies widely, not only due to differences in algorithms but also because researchers often use incompatible datasets, inconsistent labeling standards, and differing definitions of what constitutes "anomalous" behavior. Consequently, claims of high or low accuracy are difficult to generalize to real world mobile proctoring contexts. Several authors highlight that until standardized, publicly available, and ecologically valid datasets are developed, cross-study comparisons will remain unreliable, and meaningful benchmarking of mobile proctoring technologies will be extremely challenging [4 - 5], [9] [41].

4.4 Ethical, privacy and fairness considerations

A recurring concern in the literature is that automated proctoring systems carry significant ethical risks, including false accusations, intrusive monitoring, and disproportionate surveillance of certain student groups. Several studies caution that algorithmic decisions, particularly those based on imperfect face or behavior detection, can misinterpret innocuous actions as cheating, producing false positives that unfairly penalize examinees and undermine trust in digital assessment systems [4 - 5]. Conversely, false negatives may allow actual misconduct to go undetected, raising questions about the reliability of fully automated exam integrity tools [9]. Lightweight on device processing approaches, such as those enabled by Haar cascade classifiers, have been proposed to mitigate privacy risks. By running locally on smartphones without continuous video transmission to external servers, these models minimize the collection and storage of sensitive biometric data and reduce exposure to breaches or unauthorized third-party access [4], [43]. However, computational efficiency alone does not guarantee ethical compliance. Even with on-device inference, systems must implement data minimization, obtain clear and informed consent, provide mechanisms for human review of flagged events, and avoid opaque decision making pipelines that leave students without recourse [5], [45]. Another major concern highlighted across multiple studies is bias in face detection and recognition algorithms, including both Haar cascades and CNN-based detectors. Detection accuracy has been shown to vary across skin tones, facial structures, hairstyles, and accessories such as glasses or head coverings, leading to systematically lower performance for certain demographic groups [9], [5], [44]. These disparities, consistent with broader findings in computer vision research, can exacerbate equity concerns in high-stakes examinations. Students with darker skin tones, non-Western facial features, or culturally specific attire may experience higher rates of false flags or missed detections, reinforcing perceptions of unfairness and discrimination [4]. Together, these findings emphasize that while technical solutions like lightweight models can enhance privacy and reduce data exposure, they do not eliminate the deeper ethical and fairness challenges inherent in automated proctoring. Ensuring mobile-based AI proctoring systems are not only efficient but also fair, transparent, and accountable requires robust safeguards, human oversight, and targeted bias mitigation strategies [43 - 45].

4.5 Identified Research Gaps

Despite increasing interest in AI-powered mobile proctoring, several critical gaps remain in the literature. First, there is a notable shortage of standardized, mobile focused datasets that accurately reflect real world examination conditions, including diverse lighting, variable camera angles,

heterogeneous smartphone models, and socio-cultural differences. Most current studies rely on laboratory datasets or synthetic scenarios, which limits the ecological validity and generalizability of model performance [4], [9].

Second, evaluation protocols across studies remain inconsistent. Metrics such as accuracy, precision, recall, latency, and privacy indicators are reported in varying ways, making cross-study comparisons challenging and hindering the establishment of realistic expectations for mobile deployment [5].

Third, while lightweight and hybrid architectures, such as Haar cascade prefiltering combined with CNN verification, show promise for mobile efficiency, there is a lack of systematic studies that quantify the trade-offs between accuracy, latency, and energy consumption across different device types [4]. Similarly, compressed CNNs and TinyML models have been explored in isolation, but there is a scarcity of research integrating these approaches into fully mobile, on-device pipelines that preserve high accuracy without cloud dependency [9]. Fourth, while multi-modal systems improve detection reliability, privacy preserving designs remain underdeveloped, with few studies implementing differential privacy, encrypted on-device fusion, or other mechanisms to mitigate data exposure risks [5].

Fifth, robustness and fairness testing is insufficient. Studies report biases in face detection and gaze estimation across skin tones, facial features, and accessories, yet few evaluations systematically assess demographic or device related disparities [4]. Finally, field trials and real world deployments are rare, with most research conducted in controlled lab settings. This gap limits understanding of ecological validity, user experience, and ethical or legal feasibility in operational mobile examination contexts [9], [5]. Collectively, these gaps highlight the need for holistic, ethically-informed, mobile first research that integrates robust datasets, standardized evaluation, hybrid and compressed models, multi-modal privacy preservation, fairness testing, and real-world validation.

5. Conclusion

Machine learning methods for anomaly detection in online proctoring have advanced significantly. Deep learning models and multi-modal fusion approaches achieve the highest accuracy, particularly under controlled conditions. Haar cascade classifiers remain valuable for mobile deployments due to their computational efficiency and well established implementations, but they are less robust than deep models when faced with variable lighting, unusual poses, or occlusions. Hybrid pipelines and model compression techniques offer a promising balance between accuracy and the constraints of mobile devices. However, the field is still limited by a lack of standardized mobile-focused datasets and few real world deployments, making it difficult to confidently assess the readiness of these systems for high stakes examinations.

5.1 Future Directions for Research

Future research should prioritize the development of mobile centric benchmark datasets that reflect real exam conditions, including variations in lighting, camera angles, smartphone models, and socio cultural contexts, with carefully annotated anomalies [4], [9]. Closely tied to this is the need for standardized evaluation protocols that define task-specific metrics, such as latency, energy per inference, and privacy impact; to enable fair and transparent comparisons across models and platforms [5]. From a methodological perspective, several studies highlight the potential of hybrid Haar CNN pipelines, where Haar cascades perform low cost, continuous monitoring and lightweight CNNs handle verification of ambiguous events. This approach warrants systematic investigation to understand the trade-offs between speed and accuracy across different mobile device classes.

Advances in model compression and TinyML, including pruning, quantization, and distillation, are also essential to support accurate, fully on-device inference without reliance on cloud processing. At the same time, multi modal, privacy preserving fusion approaches, such as encrypted or on-device integration of video, audio, and telemetry using differential privacy, offer a pathway to enhance detection reliability while minimizing data exposure risks. Equally important is the need for robustness and fairness testing, including demographic bias evaluations and adversarial stress testing across device types, to ensure equitable model behavior.

Finally, researchers consistently stress the importance of real world pilot studies with human oversight to assess ecological validity, user experience, and regulatory compliance before large-scale deployment in high stakes mobile examinations.

Funding

This research received no external funding.

Data Availability Statement

Data sharing is not applicable to this article as no datasets were generated or analyzed.

Ethical considerations

The authors strictly adhered to the research ethical policies as set out in the Journal and plagiarism prevention, use of AI, Mount Kenya University Research ethics committee guidelines and the National Research ethics policies set out in the Kenya's National Commission for Science, Technology and Innovation (NACOSTI).

Acknowledgements

The researcher acknowledges the support given by Mount Kenya University, School of Computing and Informatics, by providing the foundation for conducting this study. Appreciation is also extended to my supervisors, Dr. John Kamau, and Prof. Raymond Ongus, for their assistance, thoughtful guidance, and commitment throughout this research.

Conflict of Interest

Authors declared "No conflict of Interest"

CRediT authorship contribution statement

Conceptualization, OB, JK and RO; methodology, OB; software, OB; validation, OB, JK and RO, formal analysis, JK; investigation, OB; resources, OB; data curation, OB; writing original draft preparation, OB; writing review and editing, JK and RO; visualization, OB; supervision, JK and RO; project administration, OB; funding acquisition, OB. All authors have read and agreed to the published version of the manuscript.

References

[1] P. Bawa, "Retention in online courses: Exploring issues and solutions - A literature review", *SAGE Open*, Vol. 6, No. 1, pp. 1 – 11, 2016. https://doi.org/10.1177/2158244015621777

[2] A. W. Bates, "Teaching in a digital age: Guidelines for designing teaching and learning", *Tony Bates Associates Ltd*, 2015. https://doi.org/10.14288/1.0224023

[3] G. Nithya, P. Venkadesh, S. V. Divya, D. Philip Daniel, V. Praveenkumar and V. Suryakumar, "Multi Modal Deep Learning Framework for Malpractice Detection in Virtual Exams Using YOLOV5", *2025 8th International Conference on Circuit, Power & Computing Technologies (ICCPCT)*, Kollam, India, pp. 624 - 629, 2025. https://doi.org/10.1109/ICCPCT65132.2025.11176713

[4] N. D. Lane *et al*., "DeepX: A Software Accelerator for Low-Power Deep Learning Inference on Mobile Devices", *2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, Vienna, Austria, pp. 1 - 12, 2016. https://doi.org/10.1109/IPSN.2016.7460664

[5] J. Deng, W. Dong, R. Socher, L. J. Li, Kai Li and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database", *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, pp. 248 - 255, 2009. https://doi.org/10.1109/CVPR.2009.5206848

[6] T. L. Dang, N. M. N. Hoang, T. V. Nguyen, H. V. Nguyen, Q. M. Dang, Q. H. Tran, and H. H. Pham, "Auto-proctoring using computer vision in MOOCs system", *Multimedia Tools and Applications*, Vol. 84, No. 23, pp. 26187 - 26213, 2025. https://doi.org/10.1007/s11042-024-20099-w

[7] D. Wang, Z. Shen, J. Shao, W. Zhang, X. Xue and Z. Zhang, "Multiple Granularity Descriptors for Fine Grained Categorization", *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, pp. 2399 - 2406, 2015. https://doi.org/10.1109/ICCV.2015.276

[8] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", *Proceedings of the 2001 IEEE Computer Society*

Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, pp. I- I, 2001. https://doi.org/10.1109/CVPR.2001.990517

[9] Z. Zhang, "A flexible new technique for camera calibration", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 11, pp. 1330 - 1334, Nov 2000. https://doi.org/10.1109/34.888718

[10] T. Shen, M. Zhao and S. Zhang, "Classification of Violations in the Examination Room Based on Deep Detection Algorithm", *2023 International Conference on Computers, Information Processing and Advanced Education (CIPAE)*, Ottawa, ON, Canada, pp. 100 - 107, 2023. https://doi.org/10.1109/CIPAE60493.2023.00025

[11] V. Hariprasad, A. F. Fernandez, P. N and T. N, "Exam Proctoring System Using Machine Learning", *2024 International Conference on Smart Technologies for Sustainable Development Goals (ICSTSDG)*, Chennai, India, pp. 1 - 7, 2024. https://doi.org/10.1109/ICSTSDG61998.2024.11026486

[12] I. Damaj and J. Yousafzai, "Simple and accurate student outcomes assessment: A unified approach using senior computer engineering design experiences", *2016 IEEE Global Engineering Education Conference (EDUCON)*, Abu Dhabi, United Arab Emirates, pp. 204 - 211, 2016. https://doi.org/10.1109/EDUCON.2016.7474554

[13] H. Zhang and M. A. Babar, "An Empirical Investigation of Systematic Reviews in Software Engineering", *2011 International Symposium on Empirical Software Engineering and Measurement*, Banff, AB, Canada, pp. 87 - 96, 2011. https://doi.org/10.1109/ESEM.2011.17

[14] W. Ahmed, V. K. Kommineni, B. König-Ries, J. Gaikwad, L. Gadelha, and S. Samuel, "Evaluating the method reproducibility of deep learning models in biodiversity research", *PeerJ Computer Science*, Vol. 11, art. no. e2618, 2025. https://peerj.com/articles/cs-2618/

[15] J. S. Paul, O. Farhath and M. P. Selvan, "AI based Proctoring System – A Review", *2024 International Conference on Inventive Computation Technologies (ICICT)*, Lalitpur, Nepal, pp. 1 - 5, 2024. https://doi.org/10.1109/ICICT60155.2024.10544779

[16] M. Ramzan, A. Abid, M. Bilal, K. M. Aamir, S. A. Memon and T. S. Chung, "Effectiveness of Pre-Trained CNN Networks for Detecting Abnormal Activities in Online Exams", *IEEE Access*, Vol. 12, pp. 21503 - 21519, 2024. https://doi.org/10.1109/ACCESS.2024.3359689

[17] Y. Jin, R. Zhong, S. Long and J. Zhai, "Efficient Inference for Pruned CNN Models on Mobile Devices with Holistic Sparsity Alignment", *IEEE Transactions on Parallel and Distributed Systems*, Vol. 35, No. 11, pp. 2208 - 2223, 2024. https://doi.org/10.1109/TPDS.2024.3462092

[18] A. Sharifara, M. S. Mohd Rahim and Y. Anisi, "A general review of human face detection including a study of neural networks and Haar feature-based cascade classifier in face detection", *2014 International Symposium on Biometrics and Security Technologies (ISBAST)*, Kuala Lumpur, Malaysia, pp. 73 - 78, 2014. https://doi.org/10.1109/ISBAST.2014.7013097

[19] Y. Zhang, M. Zhao, L. Yan, T. Gao and J. Chen, "CNN-Based Anomaly Detection for Face Presentation Attack Detection with Multi-Channel Images", *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, Macau, China, pp. 189 - 192, 2020. https://doi.org/10.1109/VCIP49819.2020.9301818

[20] M. Kumar and K. Sharma, "Enhanced Security in Matter-Enabled Iot Networks Through Anomaly Detection", *2025 International Conference on Pervasive Computational Technologies (ICPCT)*, Greater Noida, India, pp. 496 - 500, 2025. https://doi.org/10.1109/ICPCT64145.2025.10940459

[21] Y. Li *et al*., "Real-Time Gaze Tracking via Head-Eye Cues on Head Mounted Devices", *IEEE Transactions on Mobile Computing*, Vol. 23, No. 12, pp. 13292 - 13309, Dec 2024. https://doi.org/10.1109/TMC.2024.3425928

[22] Y. Taigman, M. Yang, M. Ranzato and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification", *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 1701 - 1708, 2014. https://doi.org/10.1109/CVPR.2014.220

[23] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, pp. 886 - 893, 2005. https://doi.org/10.1109/CVPR.2005.177

[24] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection", *Proceedings. International Conference on Image Processing*, Rochester, NY, USA, pp. I - I, 2002. https://doi.org/10.1109/ICIP.2002.1038171

[25] S. Teerapittayanon, B. McDanel and H. T. Kung, "Distributed Deep Neural Networks Over the Cloud, the Edge and End Devices", *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, Atlanta, GA, USA, pp. 328 - 339, 2017. https://doi.org/10.1109/ICDCS.2017.226

[26] F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering", *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 815 - 823, 2015. https://doi.org/10.1109/CVPR.2015.7298682

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks", *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 4510 - 4520, 2018. https://doi.org/10.1109/CVPR.2018.00474

[28] Y. Akbulut, A. Şengür, Ü. Budak and S. Ekici, "Deep learning based face liveness detection in videos", *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, Malatya, Turkey, pp. 1 - 4, 2017. https://doi.org/10.1109/IDAP.2017.8090202

[29] H. W. Hsu, T. Y. Wu, W. H. Wong and C. Y. Lee, "Correlation-Based Face Detection for Recognizing Faces in Videos", *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, pp. 3101 - 3105, 2018. https://doi.org/10.1109/ICASSP.2018.8461485

[30] R. Koshy and A. Mahmood, "Enhanced Anisotropic Diffusion-based CNN-LSTM Architecture for Video Face Liveness Detection", *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Miami, FL, USA, pp. 422 - 425, 2020. https://doi.org/10.1109/ICMLA51294.2020.00074

[31] Z. T. Hossain, P. Roy, R. Nasir, S. Nawsheen and M. I. Hossain, "Automated Online Exam Proctoring System Using Computer Vision and Hybrid ML Classifier", *2021 IEEE International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things (RAAICON)*, Dhaka, Bangladesh, pp. 14 - 17, 2021. https://doi.org/10.1109/RAAICON54709.2021.9929456

[32] T. Fatima, F. Azam and A. W. Muzaffar, "A Systematic Review on Fully Automated Online Exam Proctoring Approaches," *2022 24th International Multitopic Conference (INMIC)*, Islamabad, Pakistan, pp. 1 - 5, 2022. https://doi.org/10.1109/INMIC56986.2022.9972964

[33] Y. Atoum, L. Chen, A. X. Liu, S. D. H. Hsu and X. Liu, "Automated Online Exam Proctoring", *IEEE Transactions on Multimedia*, Vol. 19, No. 7, pp. 1609-1624, 2017. https://doi.org/10.1109/TMM.2017.2656064

[34] M. Labayen, R. Vea, J. Flórez, N. Aginako and B. Sierra, "Online Student Authentication and Proctoring System Based on Multimodal Biometrics Technology", *IEEE Access*, Vol. 9, pp. 72398 - 72411, 2021. https://doi.org/10.1109/ACCESS.2021.3079375

[35] A. A. Turani, J. H. Alkhateeb and A. A. Alsewari, "Students Online Exam Proctoring: A Case Study Using 360 Degree Security Cameras", *2020 Emerging Technology in Computing, Communication and Electronics (ETCCE)*, Bangladesh, pp. 1 - 5, 2020. https://doi.org/10.1109/ETCCE51779.2020.9350872

[36] J. H. Y. Ho *et al.*, "IoT-Enhanced Remote Proctoring: A New Paradigm for Remote Assessment Integrity", *2023 IEEE 35th International Conference on Software Engineering Education and Training*, Tokyo, Japan, pp. 197 - 198, 2023. https://doi.org/10.1109/CSEET58097.2023.00045

[37] W. Yaqub, M. Mohanty and B. Suleiman, "Privacy-Preserving Online Proctoring using Image-Hashing Anomaly Detection", *2022*

International Wireless Communications and Mobile Computing (IWCMC), Dubrovnik, Croatia, pp. 1113 - 1118, 2022. https://doi.org/10.1109/IWCMC55113.2022.9825119

[38] E. A. Alkinani, "Multimodal Transformer Framework for Real-Time Cheating Detection in Online Assessments and E-learning Platforms", *SN Computer Science*, Vol. 7, No. 1, art. no. 101, 2025. https://doi.org/10.1007/s42979-025-04718-3

[39] H. Li, M. Xu, Y. Wang, H. Wei, and H. Qu, "A visual analytics approach to facilitate the proctoring of online exams", *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1 - 17, 2021. https://doi.org/10.1145/3411764.3445294

[40] X. Wang, N. Su, Z. He, Y. Liu, and S. Ma, "A large-scale study of mobile search examination behavior", *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1129 - 1132, 2018. https://doi.org/10.1145/3209978.3210099

[41] C. Mutimukwe, O. Viberg, C. McGrath, and T. Cerratto-Pargman, "Privacy in online proctoring systems in higher education: stakeholders' perceptions, awareness and responsibility", *Journal of Computing in Higher Education*, pp. 1 - 30, 2025. https://doi.org/10.1007/s12528-025-09461-5

[42] M. Chougule, S. Bagul, M. Gharat, S. Malve and D. Kayande, "ProctoXpert – An AI Based Online Proctoring System", *2024 3rd International Conference for Innovation in Technology (INOCON)*, Bangalore, India, pp. 1 - 8, 2024. https://doi.org/10.1109/INOCON60754.2024.10511868

[43] N. Y. Al Hoqani, T. Regula, S. G. K. Kumar, M. R. Battina, A. N. Al Salmi and K. Ayyaraju, "Artificial Intelligence-Enabled Online Exam Proctoring Systems: Technologies, Challenges, and Scalable Solutions for Remote Education," *2025 IEEE International Conference on Computation, Big-Data and Engineering (ICCBE)*, Penang, Malaysia, pp. 179 - 183, 2025. https://doi.org/10.1109/ICCBE65177.2025.11255752

[44] Y. Singh, R. R. Nair, T. Babu, and P. Duraisamy, "Enhancing academic integrity in online assessments: Introducing an effective online exam proctoring model using yolo", *Procedia Computer Science*, Vol. 235, pp. 1399 - 1408, 2024. https://doi.org/10.1016/j.procs.2024.04.131

[45] N. Malhotra, R. Suri, P. Verma and R. Kumar, "Smart Artificial Intelligence Based Online Proctoring System", *2022 IEEE Delhi Section Conference*, New Delhi, India, pp. 1 - 5, 2022. https://doi.org/10.1109/DELCON54057.2022.9753313

[46] A. Tweissi, W. A. Etaiwi, and D. A. Eisawi, "The accuracy of AI-based automatic proctoring in online exams", *Electronic Journal of e-Learning*, Vol. 20, No. 4, pp. 419 - 435, 2022. https://doi.org/10.34190/ejel.20.4.2600

***